



Towards better privacy preservation by detecting personal events in photos shared within online social networks

Eliana Raad

► To cite this version:

Eliana Raad. Towards better privacy preservation by detecting personal events in photos shared within online social networks. Other [cs.OH]. Université de Bourgogne, 2015. English. NNT : 2015DI-JOS079 . tel-01719723

HAL Id: tel-01719723

<https://theses.hal.science/tel-01719723>

Submitted on 28 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SPIM

Thèse de Doctorat



école doctorale sciences pour l'ingénieur et microtechniques
U N I V E R S I T É D E B O U R G O G N E

Towards better privacy preservation by detecting personal events in photos shared within online social networks

ELIANA RAAD

Le 4 décembre 2015
Devant le jury composé de

Harald KOSCH
David GROSS-AMBLARD
Ernesto DAMIANI
Philippe ANIORTE
Kokou YETONGNON
Richard CHBEIR

Professeur, Université de Passau
Professeur, Université de Rennes 1
Professeur, Université de Milan
Professeur, Université de Pau et des Pays de l'Adour
Professeur, Université de Bourgogne
Professeur, Université de Pau et des Pays de l'Adour

Rapporteur
Rapporteur
Examineur
Examineur
Examineur
Directeur de thèse

**Towards better privacy preservation
by detecting personal events in
photos shared within online social
networks**

ELIANA RAAD

À mes parents

*« Partout où l'homme apporte son travail,
Il laisse aussi quelque chose de son cœur. »*

Henryk Sienkiewicz

Remerciements

J'adresse mes sincères remerciements au professeur Harald KOSCH et au professeur David GROSS-AMBLARD pour avoir consacré leur temps pour rapporter mon travail de thèse. Je remercie également professeur Ernesto DAMIANI, professeur Kokou YETONGNON et professeur Philippe ANIORTÉ d'avoir bien voulu faire partie des membres de mon jury. Merci à Philippe pour sa bienveillance, ses conseils et aussi son amabilité.

La thèse est un voyage avec un problème à résoudre certes, mais aussi une aventure que l'on vit avec des personnes pendant des années... des années de ce qui fut la période la plus intense de ma vie.

Merci au professeur Richard CHBEIR pour cette belle aventure dont il était le guide. Il a été celui qui n'a jamais cessé d'accroître ma capacité de produire et en même temps celui qui fut capable de me tirer vers le haut quand j'en avais besoin. Que ce travail soit le gage de mon infinie gratitude.

Merci au docteur Bechara AL BOUNA pour son aide depuis le début de mon travail de recherche et son soutien tant sur le plan humain que professionnel.

Merci au docteur Joe TEKLI pour ses conseils qui ont été d'une aide précieuse pour moi.

Merci au docteur Gilbert TEKLI pour ses encouragements.

Merci à toute l'équipe de l'IUT de Bayonne et le laboratoire LIUPPA pour l'accueil spontané et pour l'ambiance conviviale ainsi que leur sympathie.

Merci à Dounia RADI pour son aide et sa disponibilité.

J'ai rencontré pendant la thèse ceux qui sont devenus depuis mes meilleurs amis.

J'ai apprécié tout ce qu'on a de commun, mais aussi tout ce qu'on a de différent. L'échange culturel que j'ai vécu avec eux, entre la Chine, le Pérou, la Tunisie, l'Ethiopie, la Thaïlande, le Venezuela, l'Argentine, l'Espagne et le Liban, je ne l'aurai jamais vécu ailleurs.

Merci à Kouki, Ghada, Nathalie, Manal, Rana et Regina pour leur disponibilité dans les moments les plus délicats.

Merci à Fabien pour son écoute, sa patience, sa compréhension et son support inestimable.

Merci à mes parents qui m'ont donné l'exemple de comment aimer sans mesure, donner sans raisons et se soucier sans attentes. Ils m'ont appris l'esprit du réel amour.

Merci à Elie et Charbel, mes deux frères d'être toujours là pour moi. Merci à Mylène ma belle-sœur pour son amitié et son aide. Merci à Sylvie pour son assistance et son amabilité.

À toutes ces personnes qui ne m'ont apporté que du bonheur, merci de faire partie de ma vie!

Abstract

Today, social networking has considerably changed why people are taking pictures all the time everywhere they go. More than 500 million photos are uploaded and shared every day, along with more than 200 hours of videos every minute. More particularly, with the ubiquity of smartphones, social network users are now taking photos of events in their lives, travels, experiences, etc. and instantly uploading them online. Such public data sharing puts at risk the users' privacy and expose them to a surveillance that is growing at a very rapid rate. Furthermore, new techniques are used today to extract publicly shared data and combine it with other data in ways never before thought possible. However, social networks users do not realize the wealth of information gathered from image data and which could be used to track all their activities at every moment (e.g., the case of cyberstalking). Therefore, in many situations (such as politics, fraud fighting and cultural critics, etc.), it becomes extremely hard to maintain individuals' anonymity when the authors of the published data need to remain anonymous.

Thus, the aim of this work is to provide a privacy-preserving constraint (*de-linkability*) to bound the amount of information that can be used to re-identify individuals using online profile information. Firstly, we provide a framework able to quantify the re-identification threat and sanitize multimedia documents to be published and shared. Secondly, we propose a new approach to enrich the profile information of the individuals to protect. Therefore, we exploit personal events in the individuals' own posts as well as those shared by their friends/contacts. Specifically, our approach is able to detect and link users' elementary events using photos (and related metadata) shared within their online social networks. A prototype has been implemented and several experiments have been conducted in this work to validate our different contributions.

Résumé

De nos jours, les réseaux sociaux ont considérablement changé la façon dont les personnes prennent des photos qu'importe le lieu, le moment, le contexte. Plus que 500 millions de photos sont partagées chaque jour sur les réseaux sociaux, auxquelles on peut ajouter les 200 millions de vidéos échangées en ligne chaque minute. Plus particulièrement, avec la démocratisation des smartphones, les utilisateurs de réseaux sociaux partagent instantanément les photos qu'ils prennent lors des divers événements de leur vie, leurs voyages, leurs aventures, etc. Partager ce type de données présente un danger pour la vie privée des utilisateurs et les expose ensuite à une surveillance grandissante. Ajouté à cela, aujourd'hui de nouvelles techniques permettent de combiner les données provenant de plusieurs sources entre elles de façon jamais possible auparavant. Cependant, la plupart des utilisateurs des réseaux sociaux ne se rendent même pas compte de la quantité incroyable de données très personnelles que les photos peuvent renfermer sur eux et sur leurs activités (par exemple, le cas du cyberharcèlement). Cela peut encore rendre plus difficile la possibilité de garder l'anonymat sur Internet dans de nombreuses situations où une certaine discrétion est essentielle (politique, lutte contre la fraude, critiques diverses, etc.).

Ainsi, le but de ce travail est de fournir une mesure de protection de la vie privée, visant à identifier la quantité d'information qui permettrait de ré-identifier une personne en utilisant ses informations personnelles accessibles en ligne. Premièrement, nous fournissons un framework capable de mesurer le risque éventuel de ré-identification des personnes et d'assainir les documents multimédias destinés à être publiés et partagés. Deuxièmement, nous proposons une nouvelle approche pour enrichir le profil de l'utilisateur dont on souhaite préserver l'anonymat. Pour cela, nous exploitons les événements personnels à partir des publications des utilisateurs et celles partagées par leurs contacts sur leur réseau social. Plus précisément, notre approche permet de détecter et lier les événements élémentaires des personnes en utilisant les photos (et leurs métadonnées) partagées au sein de leur réseau social. Nous décrivons les expérimentations que nous avons menées sur des jeux de données réelles et synthétiques. Les résultats montrent l'efficacité de nos différentes contributions.

Contents

Chapter 1	Introduction	1
I.	Introduction	2
I.1	The Mass Surveillance on the Internet	3
I.2	The Issue of Privacy Protection	4
II.	Thesis Context	5
II.1	Popularity and Usage of Online Social Networks	5
II.2	Emergence of Online Sharing of Multimedia Content	7
II.3	Tendency of Sharing Photos of Personal Events	9
III.	Anonymity in Online Social Networks	9
III.1	The case of Cyberstalking	10
III.2	Motivating Scenario	11
III.3	Challenges	16
III.4	Privacy Preservation Solutions in Multimedia Publication	18
IV.	Contributions	20
IV.1	A secured publication process of multimedia documents	20
IV.2	Event detection from images shared on social networks	21
IV.3	Computing links between events on social networks	21
V.	Report Organization	21
Chapter 2	Related Works	25
	Abstract	25
I.	Introduction	26
II.	Inference Control	27
II.1	Database Inference	27
II.2	Ontology-based Inference	29
II.3	Web-based Inference	30
II.4	Discussion	32
III.	Online Identity	35
III.1	Online Identity Terminologies	35
III.2	Statistical Approaches	37

III.3	Learning-based Approaches	42
III.4	Ontological Approaches	45
III.5	Discussion	53
IV.	Events in Online Social Networks	55
IV.1	Event Definitions	55
IV.2	Event Detection from Photos in Online Social Networks	58
IV.3	Events Linking Approaches	63
V.	Conclusion	71
Chapter 3	Privacy-Preserving Framework for Multimedia Sharing	73
Abstract		73
I.	Introduction	74
II.	Adversary Model	75
III.	Data Model	76
III.1	Data Definition	76
III.2	Data Comparison	81
III.3	Identity Anonymization Problem	86
IV.	Privacy Preservation Prior to Publication	87
IV.1	Achieving de-linkability	88
IV.2	Multimedia Document Sanitization: MD^* – algorithm	89
IV.3	Utility Estimation	91
V.	Experiments	93
V.1	Dataset configuration	93
V.2	Evaluation Results	94
VI.	Conclusion	98
Chapter 4	Personal Event Detection from OSN Photo's Metadata	101
Abstract		101
I.	Introduction	102
II.	Data Model	103

II.1	User (<i>u</i>)	103
II.2	Star Social Network (<i>SSN</i>)	104
II.3	Social Space (Δ)	104
II.4	Metadata (<i>meta</i>)	105
II.5	Image (<i>img</i>)	106
II.6	Elementary Event (<i>e</i>)	107
III.	Trip Scenario	109
III.1	The multi-* property of events	109
III.2	Challenges	112
IV.	Methodology Overview	113
V.	Metadata Pre-processing	114
V.1	Time Granularity	114
V.2	Space Granularity	115
V.3	People-tags Granularity	115
VI.	Elementary Event Detection	115
VII.	Prototype and Experimentations	116
VII.1	Prototype System	117
VII.2	The MediaEval Dataset	119
VII.3	Evaluation Measurements	120
VII.4	Experimental Evaluation	123
VIII.	Conclusion	133
Chapter 5	Linking Elementary Events	135
	Abstract	135
I.	Introduction	136
II.	Definitions	139
II.1	Event Context: Who, Where, When	139
II.2	Social-R Space	139
III.	Meta-model of Event Relations	141

III.1	Why a Meta-model? _____	142
III.2	Topological Models _____	142
III.3	The 4-Intersection Model _____	143
IV.	Basic Relations between Events _____	145
IV.1	Spatial Relations _____	146
IV.2	Temporal Relations _____	149
IV.3	Social Relations _____	152
V.	Event Constraints _____	162
V.1	Time-based Constraints _____	163
V.2	Location-based Constraints _____	163
V.3	Social-based Constraints _____	163
V.4	Pruning Relations Combinations _____	163
VI.	Computing Links between Events _____	166
VI.1	From Intersection Set to Hamard Intersect Product _____	166
VI.2	From Hamard Intersect Product to Binary Triplet _____	167
VII.	Complex Relations of Events _____	168
VII.1	Logical Expression _____	168
VII.2	Ontological Relations _____	169
VII.3	Application-based Relations _____	174
VIII.	Conclusion _____	175
Chapter 6	Conclusion _____	179
I.	Contributions _____	180
II.	Future Works _____	182
II.1	Improving theoretical Approach _____	182
II.2	Improving Validation _____	182
II.3	New Research Directions _____	183

List of Figures

Figure 1: The percentage of Internet users in each age group who use social networking sites, over time, 2005-2013. (Source: <i>Pew Research Center's survey, 2013</i>).	6
Figure 2: The percentage of adults who use Facebook, LinkedIn, Pinterest, Instagram, and Twitter, by year from 2012 to 2014. (Source: <i>Pew Research Center's survey, 2014</i>).	6
Figure 3: A list of the most popular types of content shared on social networks. (Source: <i>IPSOS report, 2013</i>).	7
Figure 4: The number of photos uploaded and shared per day on selected platforms from 2005 to 2013. (Source: <i>KPCB estimates based on publicly disclosed company data, 2014</i>).	8
Figure 5: The Twitter account of François Fillon with the fake nickname “@fdbeauce”.	11
Figure 6: A fragment of the Wikipedia's page of François Fillon.	13
Figure 7: The profile image on “@fdbeauce” Twitter account.	13
Figure 8: Metadata of the profile image on “@fdbeauce” Twitter account.	14
Figure 9: Public information existing online (image and description) of “Château de Beaucé”, which is part of the cultural heritage of France.	14
Figure 10: The tweet of “@fdebeauce” on October 23, 2011 and the French News reporting François Fillon's visit to Japan for two days.	16
Figure 11: The framework for an inference detection system proposed in [32].	28
Figure 12: Online Social Footprint defined in [16].	35
Figure 13: Attribute list of personal information explored in [17].	38
Figure 14: System architecture presented in [21].	41
Figure 15: Overview of the approach proposed in [20] to disambiguate identity web references using Web 2.0 data.	44
Figure 16: Approach used in [19] for data aggregation.	45
Figure 17: Overview of the distributed di.me system architecture in [63].	50
Figure 18: The approach process proposed in [25].	52
Figure 19: Distinction between personal events and social events, as defined in [83].	57
Figure 20: Basic aspects of event description from [84].	58
Figure 21: Event detection framework presented in [82, 83].	59
Figure 22: Architecture diagram of event detection approach proposed in [81].	65

Figure 23: Illustration of atomic and composite events in [108].	67
Figure 24: Multimedia (metadata) exploration framework presented in [107].	69
Figure 25: Examples of typical image descriptions using our multimedia object representation. Figure 1(a) and Figure 1(b) show similar content, contained in two different multimedia documents.	78
Figure 26: Privacy violation evaluation.	95
Figure 27: Uncertainty evaluation.	96
Figure 28: Utility evaluation.	97
Figure 29: Computational cost evaluation.	98
Figure 30: The social space \square , modeled as a three-dimensional space: locations, times, and people names.	105
Figure 31: Example of a photo with its metadata and faces of tagged persons.	107
Figure 32: Representation of the event social space in a 3D graph.	109
Figure 33: Some photos shared on SSN_{u0} created by five different persons, including u_0 .	110
Figure 34: Some photos shared on SSN_{u0} created by <i>Lisa</i> and <i>Regi</i> where new participants (<i>Joseba</i> and <i>Irvin</i>) are identified.	110
Figure 35: Some photos shared on SSN_{u0} taken over the four days of the trip.	111
Figure 36: Some photos shared on SSN_{u0} taken in different cities during the trip.	112
Figure 37: Description of our approach to detect elementary events.	114
Figure 38: Architecture of our framework to detect events.	117
Figure 39: F-measure scores considering different time distributions.	125
Figure 40: NMI scores considering different time distributions.	125
Figure 41: F-measure scores considering different location distributions.	126
Figure 42: NMI scores considering different location distributions.	126
Figure 43: F-measure and NMI scores considering different uploaders distributions.	127
Figure 44: F-measure and NMI results considering different temporal information availabilities.	130
Figure 45: F-measure and NMI results considering different geographic information availabilities.	131
Figure 46: F-measure and NMI results considering different uploader information availabilities.	131

Figure 47: F-measure and NMI results with different metadata availabilities.	132
Figure 48: Time performance w.r.t. dataset size.	133
Figure 49: Representation of the social-R space in a 3D graph.	140
Figure 50: Transition from the social space (on the left) to the social-R space (on the right).	141
Figure 51: Lisa's social connections. Two social circles: Family: Elie, Bouli; Colleagues: Rich, Gil, Kouki, Naty, Solomon, and Minale.	146
Figure 52: Set of the six possible spatial relations between events (with their graphical description and their corresponding intersection matrices).	148
Figure 53: Set of the six possible temporal relations between events (with their graphical description and their corresponding 4-intersection matrices).	150
Figure 54: Illustration of event social interior and boundary.	153
Figure 55: Set of the six possible social relations between events (with their graphical description and their corresponding intersection matrices).	161
Figure 56: The Event Semantization process.	176

List of Tables

Table 1: Comparison of inference control approaches.	33
Table 2: Comparison of ontological approaches.	50
Table 3: Comparison of online identity approaches.	54
Table 4: A summary of photo-based event detection approaches.	62
Table 5: Comparison of event relations approaches.	70
Table 6: Notations used in our approach.	76
Table 7: Distribution of metadata in each dataset group.	124
Table 8: Availability of metadata in each dataset.	128
Table 9: F-value and NMI scores considering different granularities of time and space on the Setfull dataset.	129
Table 10: Spatial intersection sets of two events of the example.	148
Table 11: Temporal intersection sets of two events of the example.	151
Table 12: The binary and hexadecimal values of the three basic relations (spatial, temporal and social)	162
Table 13: The possible combinations of the three relations:	164

Chapter 1: Introduction

“The European Commission says data privacy is a basic human right, and yet Facebook and other social networks have over a billion people publicly broadcasting their personal data every day.”

-Steve Matthey

Chief operating officer

VCCPme

I. Introduction

Large scale web applications are gaining increasing interest in recent times across a range of sectors, in both large and small firms. Companies are now constantly looking at what kind of data they have and what data they need in order to maximize their market position. In the era of big data, there are concerns about data privacy, and even the potential future value of data, as expressed in the White House Counsel John Podesta's 2014 report to the President on the challenges of Big Data [1]. The main added privacy risk is that this data – from voice calls, emails and texts to uploaded pictures, video, and music – is being reused and combined with other data in ways never before thought possible. Further, the report states: “This is driving data collection to become functionally ubiquitous and permanent, allowing the *digital traces* we leave behind to be collected, analyzed, and assembled to reveal a surprising number of things about ourselves and our lives”.

With all these developments, the ever-increasing amount of information flowing through social media and blogging sites has reflected the need for heightened privacy controls. People are sharing and uploading upwards of 1.8 billion photos¹ a day². In many situations, motivated by several campaigns such as politics, fraud fighting, cultural critics, and others, authors of some of these social medias need to remain anonymous. Consequently, when a data provider outsources or publishes multimedia documents, it becomes extremely hard sometimes to maintain individuals' anonymity mainly due, but not limited to: 1) the number of active social networks to which they actually participate, and 2) the trails of seemingly information they leave behind [2]. These trails of information make individuals victims of what is known by the Internet community as *cyberstalking* where an adversary clandestinely tracks the movements of an individual.

¹ Both terms photo and image will be used interchangeably in the remainder of this study.

² <http://www.kpcb.com/internet-trends>

I.1 The Mass Surveillance on the Internet

Technology has brought many advances and conveniences in various fields of study, including public health, education and transportation. However, it also comes with the cost of privacy. Many examples have been seen in the news (e.g., the US National Security Agency - NSA³- spying on German chancellor Angela Merkel⁴, NSA spying on 98 percent of South American communications⁵, etc.). “*Everyone is under surveillance now*”⁶, says Edward Snowden⁷, who was responsible for leaking PRISM, one of the surveillance programs of the United States government. PRISM is a tool used by the NSA to collect private electronic data belonging to users of major Internet services like Gmail⁸, Facebook⁹, Outlook¹⁰, and others¹¹. NSA programs collect two kinds of data: metadata and content. In essence, Snowden’s disclosure was the entry of the term ‘*metadata*’ into common usage. This is information about the time and location of a phone call or email. In addition to written and oral communications, the NSA has collected millions of faces from web images to develop the large untapped potential of using facial images, fingerprints, and other identifiers to track suspected terrorists and other intelligence targets¹².

³ <https://www.nsa.gov/>

⁴ <http://www.theguardian.com/us-news/2015/jul/08/nsa-tapped-german-chancellery-decades-wikileaks-claims-merkel>

⁵ <http://techcrunch.com/2015/07/08/assange-nsa-intercepts-98-of-south-american-communications/>

⁶ http://www.theguardian.com/world/2014/may/03/everyone-is-under-surveillance-now-says-whistleblower-edward-snowden?CMP=twl_gu

⁷ A former employee of the CIA and of a private contractor working for the US National Security Agency (NSA)

⁸ <https://mail.google.com/>

⁹ <https://www.facebook.com/>

¹⁰ www.microsoft.com/en-us/outlook-com/

¹¹ <http://www.theverge.com/2013/7/17/4517480/nsa-spying-prism-surveillance-cheat-sheet>

¹² See James Risen and Laura Poitras, , “NSA Collecting Millions of Faces From Web Images”, in: The New York Times of 31 May 2014

On the one hand, the NSA claims that it absolutely needs all this data to help prevent terrorist attacks. On the other hand, privacy activists critical of the NSA surveillance program disagree, arguing not only that the collection is based on a legal interpretation that goes way beyond what Congress allowed, but also that metadata includes personal information, which can build a more detailed profile even than listening into content¹³.

I.2 The Issue of Privacy Protection

The problem particularly occurs when the so-called "metadata" are shared with others and combined with publicly available information. For example, in a study released in January 2015, titled "*Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata*", a group of data scientists analyzed credit card transactions made by 1.1 million people in 10,000 stores over a three-month period¹⁴. The database did not have any names, account numbers, or other obvious identifying features. Nevertheless, even with this anonymous data, data scientists could re-identify 90% of the shoppers as unique individuals and could uncover their records, just by knowing a few bits of information about them. And that uniqueness of behavior combined with publicly available information in multimedia documents, such as Instagram or Twitter posts, made it possible to re-identify people's records by name.

Personal data means any information relating to an identified or identifiable individual. An identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number (e.g., social security number) or one or more factors specific to his physical, physiological, mental, economic, cultural or social identity (e.g., name and first name, date of birth, biometrics data, fingerprints, DNA, etc.)¹⁵.

¹³ <http://www.theguardian.com/world/interactive/2013/nov/01/snowden-nsa-files-surveillance-revelations-decoded#section/2>

¹⁴ <http://dujs.dartmouth.edu/computer-science/making-data-anonymous-not-enough-to-protect-consumer-privacy#.VgqxDMuqqko>

¹⁵ <http://www.cnil.fr/english/data-protection/personal-data-definition/>

Therefore, even when all personally identifying information, such as name, birthdate, address, is completely removed from a database, that data is not as entirely anonymous as many people think. All of this bears out the need of better means to *anonymize* all multimedia documents that can be attributed to anonymous users. Unfortunately, removing all “personally identifying information” is not a trivial operation with all the digital trails users leave today in their regular activities on the Internet (e.g., blogging, social networking, and file sharing, etc.). Meanwhile, requiring an absolute guarantee that a data set cannot be linked back to individuals probably might render it useless in some cases or discourage any sharing. “*Finding the right balance between privacy and utility is absolutely crucial to realizing the great potential of metadata*”, say the researchers of the Credit Card Metadata study.

Anonymization is the modification of data so that sensitive information remains private. De-anonymization is the converse: re-identifying somebody in an anonymized network – or even simply learning something about them that was not meant to be attributable to them.

II. Thesis Context

In this work, we are interested in personal data that is left in the raw data of multimedia publications and that can be attached to an individual. In this section, we outline the main motivations behind our interest in exploiting metadata of photos shared within online social networks (OSN).

II.1 Popularity and Usage of Online Social Networks

Nowadays, the popularity of social networks is constantly growing. According to a survey conducted by the Pew Research Center¹⁶, the number of social networks’ users has shown an explosive growth between 2005 and 2013, as illustrated in Figure 1. Seventy four percent (74%) of American Internet adults use social networking sites, and even more than their half (52%) now use two or more social media sites compared with 42% who did so in 2013.

¹⁶ http://www.pewinternet.org/files/2015/01/PI_SocialMediaUpdate2014_pdf.pdf

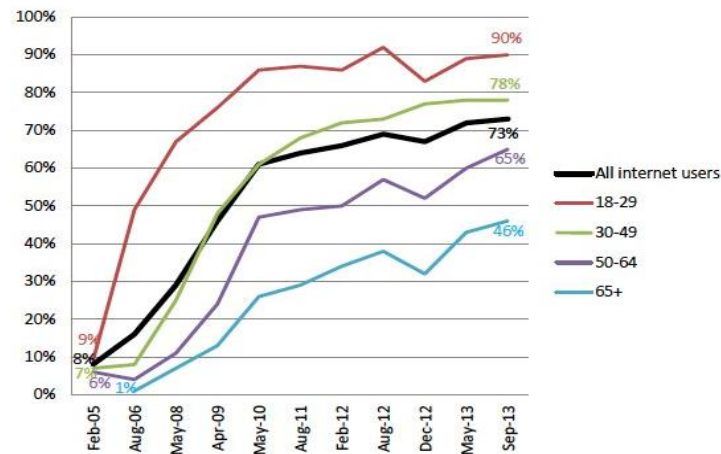


Figure 1: The percentage of Internet users in each age group who use social networking sites, over time, 2005-2013. (Source: Pew Research Center's survey, 2013).

This popularity makes social networking one of the main activities on the Web today. For instance, while Facebook remains the most popular social media site, other social media platforms like Twitter, Instagram, Pinterest and LinkedIn saw significant growth between 2012 and 2014. The results from these estimates are shown in Figure 2.

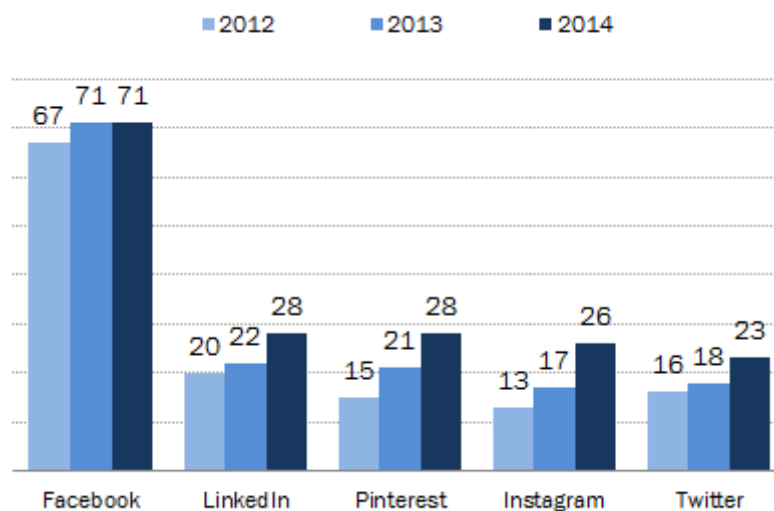


Figure 2: The percentage of adults who use Facebook, LinkedIn, Pinterest, Instagram, and Twitter, by year from 2012 to 2014. (Source: Pew Research Center's survey, 2014).

II.2 Emergence of Online Sharing of Multimedia Content

With the advent of new types of media and the diversity of technologies supporting multimedia, it became easy to take a video, capture an audio or snap a picture. Furthermore, user-friendly social networks applications made multimedia sharing easy and therefore a popular activity. According to a report on Ipsos¹⁷, a survey indicated that pictures are the most popular shared item on social media sites (43% as seen in Figure 3).

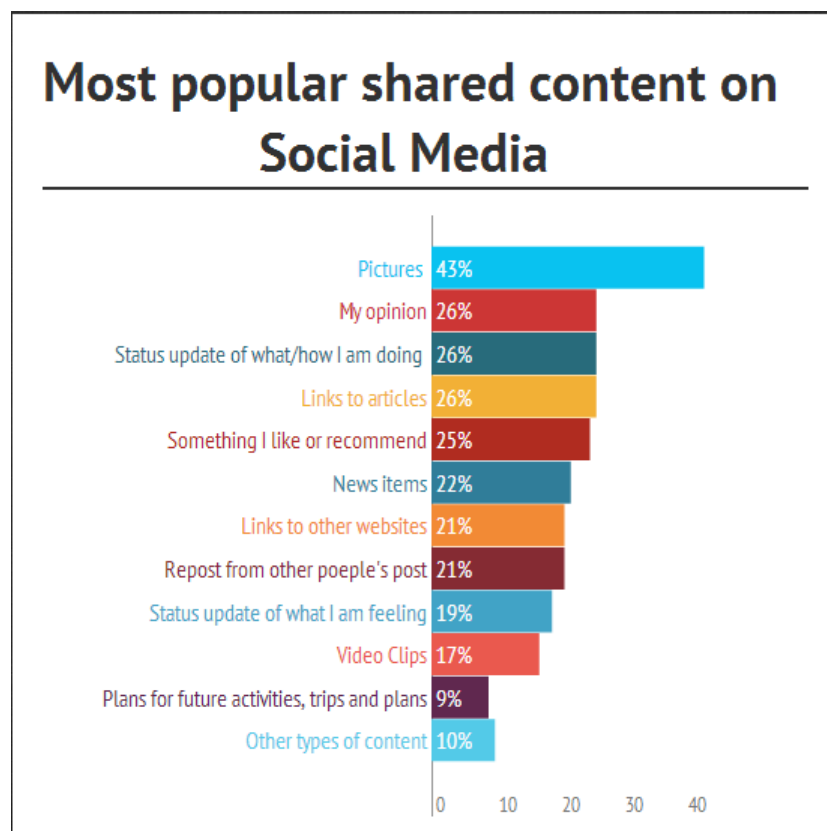


Figure 3: A list of the most popular types of content shared on social networks. (*Source: IPSOS report, 2013*).

¹⁷ <http://www.socialmediatoday.com/content/what-most-popular-content-shared-social-media>

Indeed, photos have become the primary content form of the online sharing, expressing easily a variety of information such as interests, activities and life experiences. Figure 4 illustrates the daily number of photos uploaded and shared on social media (Facebook, Flickr, Instagram, Snapchat, and WhatsApp). However, social network users do not seem to fully realize the wealth of information (metadata) that can be obtained from those photos and which could be used to track their activities. For instance, the project “*I Know Where Your Cat Lives*”¹⁸ raises concerns over online privacy by using public images of cats uploaded to photo sharing websites (e.g., Flickr¹⁹, Twitpic²⁰ and Instagram). Location coordinates embedded in the images’ metadata are used to show where each cat lives and, more importantly, to track their owner's homes.

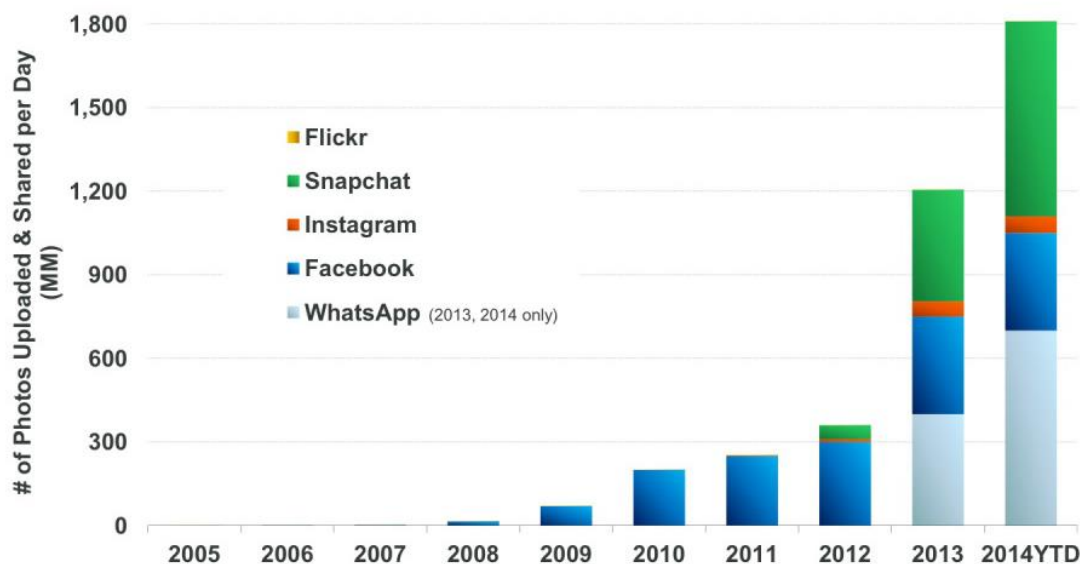


Figure 4: The number of photos uploaded and shared per day on selected platforms from 2005 to 2013. (Source: KPCB estimates based on publicly disclosed company data, 2014).

¹⁸ <http://iknowwhereyourcatlives.com/cat/e3fb32d4a0>

¹⁹ <https://www.flickr.com/>

²⁰ <http://twitpic.com/>

II.3 Tendency of Sharing Photos of Personal Events

People are increasingly sharing their daily lives online, exposing their everyday movements and putting more and more from their personal information online. From the coffee mug for *#Breakfast* to the *#Sleeping#Cat*, people are snapping photos of everything. Sharing real-life experiences with friends at any time and any place has become common on social networks, from social events (festive celebrations, concerts, sporting events, etc.) and personal events (family gathering, wedding and other parties, etc.) to the most intimate private moments, conducting sometimes to breach users' privacy and scandals.

Furthermore, people are constantly using their cell phones to capture events happening around them, whether they participated in or just observed. The evolution of smartphones has therefore changed the way people share their events and the way they interact with them as well. It is no more surprising that cameras and phones are these days present all the time in our life, even during others' precious moments such as weddings and births. People have all become so used to sharing instantly on social media applications what they are doing that they spend much time recording, for the purpose of sharing their experiences and activities online.

Because photos posted on social networks represent real stories from users' life, we focus in this work on detecting events from photos uploaded online on photo sharing websites.

III. Anonymity in Online Social Networks

Online privacy does not only concern terrorists, hackers or child pornographers. Some people are also worried about staying safe from their own governments or employers. Others might not want to let their online audience know their sex or sexuality, or just do not want their names attached to a comment or a question online. To be anonymous, means one is not identified by name or has an unknown name. A number of factors are likely to drive people to stay anonymous on the Internet. One of these factors is the fact that the Internet is forever. Whatever question people ask or opinion they share, there it is, more or less forever. In *"I Know Who You Are and I Saw What You Did: Social Networks and the Death of Privacy"* [3], Lori Andrews shows, on the one hand, the importance of anonymous postings in line with the

right to privacy of citizens. And, on the other, Andrews describes the potential of harmful anonymous postings in some extreme cases (e.g., defamation, discrimination, etc.).

Indeed, online anonymity does not only help protect fundamental rights, but also ensures the personal safety of individuals expressing their opinions (e.g., postings about politics, political figures, social institutions and services, etc.). Given its ability to facilitate the rich, diverse and far ranging exchange of ideas, anonymous use of the Internet allows protesters to deliver their message to a larger audience without the message being prejudged (e.g., revolutions of the “Arab Spring”²¹ in 2011).

However, anonymity on the Internet may magnify particular harms and facilitate wrongdoing (e.g., terrorism, violation copyright laws, identity theft, cyber predators, cyberstalking, etc.). This becomes part of the problem of anonymous cybercrime. Cybercriminals have found ways to monetize personal information found on the user profile pages of social networks. For instance, this type of personal identifiable information (PII) harvesting allows cybercriminals to obtain answers to security questions used to verify the user’s identity when attempting to log in to sensitive services such as online banking sites²².

III.1 The case of Cyberstalking

In the context of criminal activities involving human beings, a stalking crime is generally considered to be one in which an individual (“the stalker”) clandestinely tracks the movements of another individual or individuals (“the stalkee”). Cyberstalking can be understood as a form of behavior in which certain types of stalking-related activities - which in the past have occurred in physical space - are extended to the online world [4]. Internet stalkers can operate anonymously or pseudonymously while online.

²¹ https://en.wikipedia.org/wiki/Arab_Spring

²² <https://securityintelligence.com/how-cybercriminals-monetize-information-obtained-from-social-networks/>

In the next section, we focus our attention on a specific case of Internet stalking involving the previous French Prime Minister François Fillon. Actually, Fillon was using a fake account name “@fdbeauce” to remain anonymous on Twitter²³. He was able to operate online pseudonymously to track messages of his ministers and journalists publishing information about him. In spite of this, we will see how it was possible for the *adversary* to recognize his *target*, and thus break his anonymity.

An **adversary** is somebody who attempts to reveal sensitive, private information about a target. A **target** is the particular social network member against whom an adversary is trying to breach privacy.

III.2 Motivating Scenario

François Fillon²⁴, the main persona of our scenario, is a French politician present online with the Twitter account @fdbeauce²⁵ (shown in Figure 5).

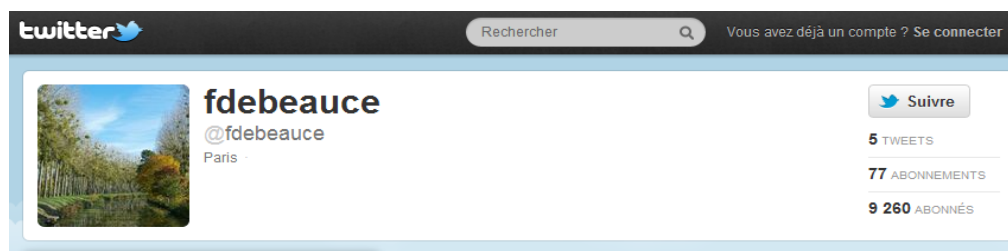


Figure 5: The Twitter account of François Fillon with the fake nickname “@fdbeauce”.

²³ <https://twitter.com/>

²⁴ François Fillon is a French lawyer and politician who served as Prime Minister of France from 2007 to 2012. https://en.wikipedia.org/wiki/Fran%C3%A7ois_Fillon

²⁵ <https://twitter.com/FrancoisFillon>

François Fillon²⁶ thought that using a fake account name on Twitter will save his privacy and keep him anonymous under the protection of the Internet security. In reality, it didn't take more than few minutes to reveal the hidden account of the French politician. In essence, the adversary used the available information on Twitter (account name, profile image and a previously published tweet) with some *background knowledge* to recognize his target.

Background knowledge is defined as any piece of information that is not directly revealed to the users but is available in the outside world or in a system and can be used to infer the attribute in question. It can also be understood as a mental model or world model that provides rules of how to link information together. Such background knowledge is often acquired from real world through experience and observation. Not only may inference require the use of background knowledge, but also the information being inferred (e.g., users' identity at physical appearance granularity) may not be stored in the application database [5].

In the following, we describe the personal data provided by François Fillon that contained enough clues resulting in *identity disclosure*.

Identity disclosure is a privacy breach in which a presumably anonymous person is in fact identifiable.

III.2.1 Profile Information

The first clue that made this attack successful is the username or the Twitter alias “@fdbeauce”. This alias is based on his real information: “F” as the initial of his name. Moreover, “debeauce” is taken from his village name “Beaucé”. A fragment of the Wikipedia's page of François Fillon is shown in Figure 6.

²⁶ <http://www.euronews.com/2011/12/12/french-pms-shy-twitter-debut/>

François Fillon

From Wikipedia, the free encyclopedia

Fillon was born in [Le Mans](#), [Sarthe](#). His father is a [civil law notary](#), while his mother, Anne Fillon, is a celebrated historian. His youngest brother, Dominique, is a talented pianist.^[3]

Fillon lives with his wife, Penelope, and five children, Marie, Charles, Antoine, Édouard and Arnaud, in the 12th century [Château de Beaucé](#), set in 20 acres (8 ha) of woodland on the banks of the [River Sarthe](#) at the famous monastery village of [Solesmes](#), near [Sablé-sur-Sarthe](#) about halfway between [Le Mans](#) and [Angers](#). M. and Mme Fillon resided in various other properties, always in the [Sarthe](#), throughout their marriage, before buying [Beaucé](#) in 1993.^[3]

Figure 6: A fragment of the Wikipedia’s page of François Fillon.

III.2.2 Metadata

The second clue is the profile image published on this account (Figure 7), and more precisely the GPS coordinates embedded in its metadata indicating that it was taken in *Beaucé*.



Figure 7: The profile image on “@fdbeauce” Twitter account.

His profile image shows a stream landscape without any description. François Fillon thought that it was therefore impossible to get any information about him from this image. However, some people were interested in exploring the metadata embedded in the image data. Metadata (see Figure 8) indicates that it was taken on *October 31, 2011* at *Beaucé*, in *Sarthe*. Fillon also shared his complete address (*postal code*, *city*, and *country*), his *phone number* and *email*, which were stored in the memory of the camera used to take the photo.

EXIF — this group of metadata is encoded in 970 bytes (0.9k)	
Exif Image Size	640 × 426
Photometric Interpretation	Color Filter Array
Make	NIKON CORPORATION
Camera Model Name	NIKON D700
Orientation	Horizontal (normal)
Software	Aperture 3.2.2
Modify Date	2011:10:31 12:55:27 1 month, 27 days, 6 hours, 49 minutes, 12 seconds ago
Artist	Francois Fillon
Copyright	Fillon 2009
Exposure Time	1/125
F Number	11.00
Exposure Program	Aperture-priority AE
ISO	200
Date/Time Original	2011:10:31 12:55:27 1 month, 27 days, 6 hours, 49 minutes, 12 seconds ago
Create Date	2011:10:31 12:55:27 1 month, 27 days, 6 hours, 49 minutes, 12 seconds ago
GPS Latitude Ref	North
GPS Latitude	47.849333 degrees
GPS Longitude Ref	West
GPS Longitude	0.249333 degrees
Resolution	72 pixels/inch

XMP — this group of metadata is encoded in 1,788 bytes (1.7k)	
XMP Toolkit	XMP Core 4.4.0
Creator	Fillon François, Fillon François
Subject	Sarthe, Beaucé, famille
Creator Work Email	fcafillon@wanadoo.fr, fcafillon@wanadoo.fr
Creator Work Telephone	01423444345, 01423444345
Creator Country	France, France
Creator City	Solesmes, Solesmes
Creator Postal Code	72300, 72300
Creator Address	Beaucé, Beaucé
Serial Number	2091810
Lens	AF-S Nikkor 50mm f/1.4G
Lens ID	202154502
Image Number	7670
Flash Compensation	0
Raw XMP data	(1,788 bytes binary data)

IPTC	
Coded Character Set	UTF8
Application Record Version	2
Keywords	Beaucé, famille, Sarthe
By-line	Fillon François, Fillon François

Figure 8: Metadata of the profile image on “@fdbeauce” Twitter account.

As shown before, François Fillon lives in a manor in *Beaucé*, in the department of Sarthe in western France. The Wikipedia page about François Fillon contains this information, including a picture of the manor where he lives, as illustrated in Figure 9. Hence, the username and profile image of François Fillon, combined with existing public knowledge from Wikipedia, allowed to re-identify him even when using a pseudonym.

ARCHITECTURE	
	Liste des réponses Affiner la recherche Autre recherche
Réponse n° 1	
	Inventaire général du patrimoine culturel édifice / site : Manoir localisation : Pays de la Loire ; Sarthe ; Solesmes aire d'étude : Sable-sur-Sarthe lieu-dit : Beaucé destinations successives : maison dénomination : manoir

Figure 9: Public information existing online (image and description) of “Château de Beaucé”, which is part of the cultural heritage of France.

III.2.3 Connections

People follow other users on Twitter to make connections. So the third clue to this attack is the connections of the previous Prime Minister. His social network on Twitter is made of French politicians since he followed some accounts of his ministers in the government who actively tweet online. One of the followers of François Fillon is Alain Lambert, a french politician. François Fillon appears in many images published by Lambert. Some of these images were taken in Beaucé (as specified in their metadata). This location information is not available from François Fillon’s profile. However, it is obtained by exploiting other users’ images in the same social network.

III.2.4 Events

The last clue that made the identification of François Fillon also possible is his first tweet “@alainlambert and I with Japanese prime minister” on *October 23, 2011*. This tweet announced a meeting with the Japanese prime minister in Japan and Alain Lambert. At the same time, the French News reported François Fillon’s visit to Japan for *two* days. We illustrate in Figure 10 the tweet of “@fdebeauce” and the French News report about François Fillon. A simple comparison between where (*Japan*), when (*October 23, 2011*) this meeting happened and who (*Japanese Prime Minister*) is involved in this event allowed to re-identify François Fillon even when using his pseudonym. Furthermore, it is interesting to note that the tweet of “@fdebeauce” and the French News report shows an event of two days. Such temporal information can be used by the adversary to learn sensitive values about his target.

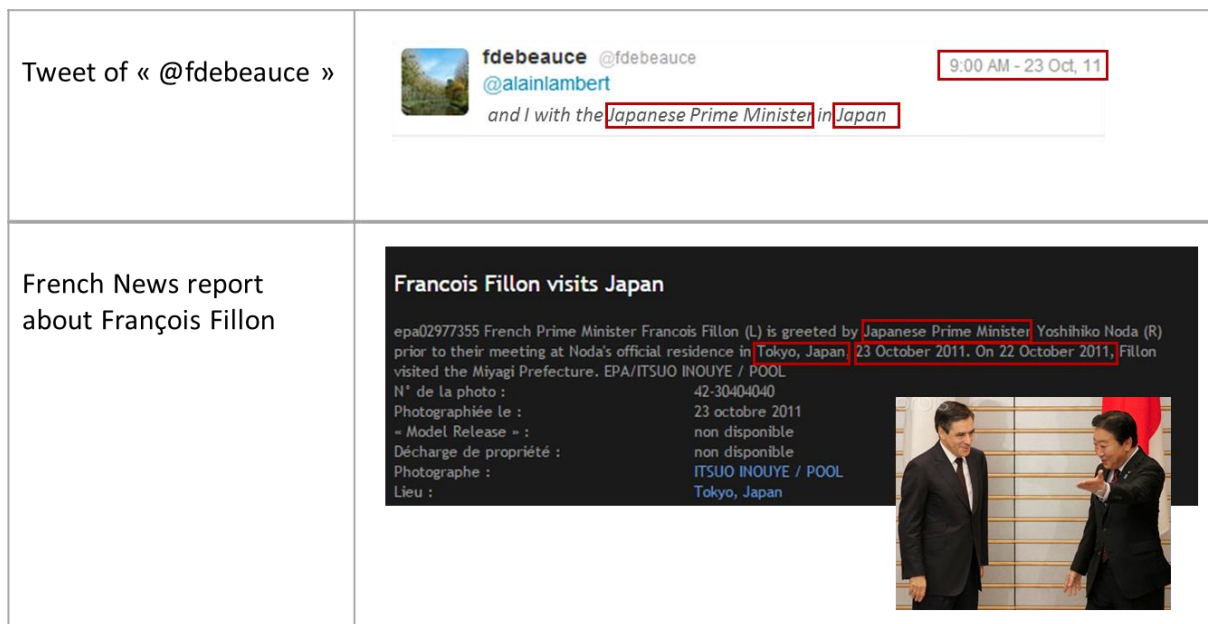


Figure 10: The tweet of "@fdebeauce" on October 23, 2011 and the French News reporting François Fillon's visit to Japan for two days.

III.3 Challenges

This scenario led us to consider three practices involving the exposure of personal information online.

First, Internet users readily provide personal data to social network sites. However, they do not realize that such data can provide further information through search engines or through combination with other users' shares.

Furthermore, users upload their photos online at any time and from anywhere. Nevertheless, they are not aware of what images can reveal about their identities. Different techniques can be applied to uploaded images. These techniques can provide extra information the users did not share, voluntarily or involuntarily, on their profiles.

Finally, users publish information about their events, travels, social activities, schedule, and so on. Initially, such data does not comprise information about their identities (e.g., the visit to

Japan in our scenario). However, the Web is continuously changing, allowing adversaries to find information about any user, and combine them with public data to infer hidden personal user information (the real name in our scenario). Moreover, adversaries are able to link information together to infer more personal user information, such as social relations, regular locations, etc.

Following this scenario, we believe that the challenges faced to prevent the identification of an anonymous user from multimedia documents are the following:

III.3.1 Remove sensitive information of both textual and multimedia content

The first challenge is to identify data (either provided by the user or propagated from the user's *connections*) in the OSN, which may contain clues about the user's *profile* and/or *events*. This data must be removed since it can be used to infer the identity of an anonymous user.

III.3.2 Consider the complex nature of *multimedia* objects

The second challenge is to protect the privacy of users publishing any multimedia document. More specifically, even if multimedia documents do not contain images of people, anonymous users can be readily identified from the combination of some metadata with other public multimedia document.

III.3.3 Preserve the *utility* of multimedia documents

The third challenge is to anonymize all sensitive data to ensure that multimedia documents shared online cannot be used to disclose the identity of the individual while keeping the utility of those shares. Let us go back to the tweet example of François Fillon: “@alainlambert and I with Japanese prime minister”. Removing all clues from the tweet of “@fdebeauce” might achieve perfect privacy, but it will be of total uselessness if the tweet results as the following: “and I with”.

III.4 Privacy Preservation Solutions in Multimedia Publication

Currently, there are many ways users can protect their privacy when publishing multimedia documents. In the following, we suggest two alternatives that users can use to stay anonymous online:

III.4.1 Burying personal data with outdated or false information

One alternative to protect online privacy is to alter personal data of the user who wants to remain hidden. This can be done by using tools that corrupt personal data existing about users online. The tool Undefined²⁷ is able to post content on behalf of user's social network account and to alter his navigation in order to confuse the components of the digital identity. Different strategies can be used to send corrupted data online, ranging from sending similar data (supporting his identity), to incomplete data (making it more confusing). Here we describe some possible ways to alter personal data online.

- Split: conflicting data is sent online,
- Disappearance: authentic data drowned in many other data,
- Prosperity: data are seconded by similar data,
- Shyness: data are minimized in their content and in their reading,
- Deception: data are sent based on previous data,
- Punctuality: Data are sent in a regular and precise way.

The following example is a real case of burying personal data with false information (i.e., images) in order to stay (visually) anonymous online. Jonathan Hirshon²⁸ asked his friends to tag him in random photos. These photos ranged from a pig pushing a shopping cart to a Buddha chill-out poster to Bill Nye the Science Guy to Don Draper to a Masonic logo, Brad Pitt, a cat, Abraham Lincoln, a seagull, and so on... Using this technique, the collection of false positives associated with the tag "Jonathan Hirshon" could confuse Facebook and

²⁷ <http://vincentdubois.fr/undefined.php>

²⁸ <http://www.fastcompany.com/3049569/has-this-man-unlocked-the-secret-to-internet-anonymity>

Google's algorithms. This “algorithm hacking” permitted to bury any real photo of Hirshon. This works well since a Google image search for his name does not return a single picture that is actually portraying him.

III.4.2 Removing or untagging content

Another alternative to preserve anonymity is to remove any published data that may be tied to a particular user. However, this is not always possible when the user is not the originator of the publication. The National Commission on Informatics and Liberty²⁹ (which is an independent administrative authority that operates in accordance with the data protection legislation, and also known as the CNIL) promoted the "right to be forgotten" and the "right to delisting". We will review briefly these data protection rights in the following:

The European Court of Justice enshrined the right to be forgotten in a 2014 judgment³⁰. Specifically, anyone who wishes to delete one or more results appearing from a search of his name can make a request to search engines, in particular Google. Google then reviews the application and grants the right if the legal conditions are met. The right to be forgotten is practically applied through the right to obtain the deletion of personal information, or the right to delisting on all of the search engine's domain names.

In practical terms, any individual who wants to see removed one or more results displayed following a search made on the basis of his/her name can make a request to a search engine, under certain conditions which take into account the public's right to information. Delisting does not lead to deletion of the information on the source website. The original information remains unchanged and will still be found on search engines using other search terms, or by direct access to the publisher's original source³¹.

²⁹ <http://www.cnil.fr/english/>

³⁰ <https://www.rudebague.com/2015/09/23/right-to-be-forgotten-cnil-and-googles-arm-wrestling-match-is-only-the-symptom-of-a-deeper-disease/>

³¹ <http://www.cnil.fr/english/news-and-events/news/article/questions-on-the-right-to-delisting/>

III.4.3 Privacy Preservation prior to Publication

Complete burying or removal of all (directly or indirectly) identifying personal data is impossible on the Internet today. Currently, new techniques are used to extract publicly shared data and combine it with other data in ways never before thought possible. No matter how much users cover their publications or posts, there could be always some information that can potentially be used to potentially identify them. Hence, controlling multimedia content prior to its publication is necessary to better prevent violation of anonymity. For this reason, we believe privacy protection (i.e., data anonymization) must be addressed prior to data sharing in online social networks.

IV. Contributions

In a viewpoint entitled "*Managing Your Digital Life*"[6], Serge Abiteboul et al. stress the need for Internet users to manage themselves their personal data. Giving users more control over how others gather and use their personal data is now more than essential. To tackle this privacy problem, we suggest in our research to let an individual control the sharing of photos made (by himself or by other users) on social media sites such as Google or Facebook. According to the authors of [6], we live in a world where it is easy to publish information but difficult to remove it or sometimes to simply access it. Therefore, users must be warned of potential violation of their privacy, before it is too late. The main contributions of this work can be summarized as follows:

IV.1 A secured publication process of multimedia documents

We provide a privacy-preserving constraint (*de-linkability*) to bound the amount of information that can be used to re-identify individuals using online profile information. *De-linkability* ensures that individuals' identifiable information composed of both textual and

multimedia content cannot be used to infer his/her identity. We formally define the identity anonymization problem in multimedia documents. Then, we provide a framework able to quantify the re-identification threat to sanitize multimedia documents to be published (*MD**-algorithm) and preserve at the same time their utility. Therefore, we provide a utility measure to determine to which extent a multimedia document remains consistent after the sanitizing process. Finally, we show a set of experiments elaborated to demonstrate the efficiency of our technique on real-world datasets.

IV.2 Event detection from images shared on social networks

We suggest a new approach to enrich the profile information of the individuals to protect. Following our motivating scenario, we exploit personal events in the individuals' own posts as well as those shared by their friends/contacts. Specifically, our approach is able to detect users' elementary events using photos (and related metadata) shared within their online social networks. We also developed a prototype tool called *Foto2Event* to validate the relevance of our clustering algorithm on large datasets.

IV.3 Computing links between events on social networks

We present a meta-model that allows representing, combining and inferring inter-event relations in an expressive and flexible way. Our meta-model allows the discovery of relations that can exist between events detected in photos shared online. Using our model, we can automatically generate relations that correspond to different aspects of events based on a homogeneous representation (spatial, temporal and social). In addition, we present a methodology that can combine spatial, temporal and social relations of the event for modeling complex relations (e.g., ontological relations such as *subevents*, *isA*, etc. or application-based relations such as *atHome*, *atAirport*, etc.).

V. Report Organization

The chapters of this report are organized as follow:

Chapter 2 gives a review related to i) inference detection, ii) construction of online identity, and iii) event detection and linking approaches.

Chapter 3 presents our privacy preservation framework. We propose *de*-linkability, a privacy-preserving constraint to bound the amount of information shared that can be used to re-identify individuals. We describe a generic data model to deal with any type of multimedia data. Then, we introduce operators necessary to achieve *de*-linkability. We also provide a sanitizing *MD* *-algorithm to enforce *de*-linkability along with a utility function to evaluate the utility of multimedia documents that is preserved after the sanitizing process. Finally, we show the set of experiments we elaborated to demonstrate the efficiency of our technique.

Chapter 4 describes our model for detecting personal events using photos' metadata shared within online social networks. We consider a specific scenario in order to describe key challenging issues regarding event detection. Then, we present a clustering algorithm able to detect and link users' elementary events using photos (and related metadata) shared within their online social networks. We present our prototype *Foto2Event* and experimental tests conducted on the MediaEval dataset.

Chapter 5 introduces our methodology for describing links between events. First, we describe our meta-model based on the 4-Intersection Model. Then, we present our methodology to identify topological relations based on three main aspects of events: spatial, temporal and social. We explain how we can combine those basic relations to infer more complex relations such as ontological relations and application-based relations.

Chapter 6 concludes our work and presents some of our future research directions.

Chapter 2: Related Works

Abstract

The focus of this work is on understanding privacy preserving techniques to prevent privacy violations via data inferences that occur when background knowledge is combined with non-private data. We first address privacy threats due to social inferences. Social inferences are the subset of inferences that results from using social applications and can pose serious threats to the privacy of their users. Such inferences are often enabled by the public sharing of personal information through social media. For this reason, we present in a second phase approaches related to the construction of online identity from multimedia data shared on online social networks. Then, we provide a review on different approaches related to events detected on online social networks. We are mainly interested in two tasks: (1) Detecting events using shared photos and (2) Linking events. In this chapter, we summarize and discuss the limitations and features of the outlined approaches.

I. Introduction

The rapid growth in the electronic exchange of personal information through social media and blogging sites today has reflected the need for privacy protection. One of the biggest issues related to online privacy is online anonymity. In many cases, users may wish to stay anonymous to protect themselves from identification. However, with the increasing volume of published data, their personally identifiable information can be readily inferred from the publicly available information and further can disclose their identities. This type of privacy breach is known as identity disclosure in the literature.

In the last decade, more and more researches have focused on privacy-preserving data mining [7-11]. The aim of privacy-preserving data mining (PPDM) algorithms is to extract relevant knowledge from large amounts of data while protecting at the same time sensitive information [12]. Different techniques have been proposed so far, which include data modification, data encryption, and data anonymization [13]. While data encryption requires a higher computing cost, data modification is widely used and includes a number of techniques such as data swapping, data shuffling, additive noise, etc. [14]. Anonymization is the modification of data so that sensitive information remains private. Data anonymization techniques such as k -anonymity has proven to be an effective means to reduce the risk of privacy disclosure via data inferences [15].

In this chapter, we present several approaches developed for data anonymization for disclosure and inference control. More specifically, we focus on privacy violations via data inferences that occur when background knowledge is combined with non-private data. This knowledge could be stored in databases, ontologies, or might be available on the Web. Most of the works done in the literature to preserve anonymity on the Web assume that identifiable information and adversaries' background knowledge are needed to make inferences. Therefore, it is of importance to select relevant information about social network users that an adversary can attribute to them. Here comes the importance of online identity.

Online identity is an important issue in today's digital world as people are becoming more open to information exposure. We review existing approaches for public online identity construction. These approaches can be classified in three categories: Statistical approaches [16-18], learning-based approaches [19-21] and ontological approaches [22-26]. Then we

compare them and identify their strength and weakness. More specifically, we point out one major drawback related to the absence of the users' personal events in their approaches.

To overcome this drawback, we also explore event detection approaches in this study to build richer profile of social network users. First, we present existing event definitions and representations. Then, we classify existing algorithms for event detection from online social networks into two classes: event clustering and event hybrid approaches. Finally, we analyze their ability to consider links between events.

In this chapter, we first present related works in the area of inference detection and privacy preservation techniques. Then, we cover online identity construction approaches. Finally, we provide a review on different approaches related to event detection and linking from multimedia data (i.e., photos) in online social networks. In each subsection, we summarize, compare and discuss the limitations and features of the outlined approaches.

II. Inference Control

The need for privacy preservation in publicly released data is not new. In essence, anonymization on relational data has been studied for decades and has inspired many approaches and methods [27-30] for social network anonymization. For this purpose, we start by presenting different approaches that have been proposed in the literature to handle identity anonymization in multimedia documents. These approaches can be categorized into three groups:

- Database inference approaches,
- Ontology-based approaches, and
- Web-based approaches.

II.1 Database Inference

Access control mechanisms are commonly used to provide control over who may access sensitive information. But in most cases, the problem is not access control, but inference

control [15, 31]. In essence, malicious users may access innocuous data, and infer sensitive information from the received answers. To address this problem, a semantic inference model (SIM) is constructed in [32, 33] based on a domain semantic knowledge. SIM represents the dependent and semantic relationships among attributes of all the entities. The inference detection system proposed in [32] consists of three modules: knowledge acquisition, semantic inference model (SIM), and security violation detection including user social relation analysis, as shown in Figure 11. It utilizes both the user's current query and past query log to determine if the current query answer can infer sensitive information. An extension of the system can be achieved by considering the cases of multiple collaborative users based on the query history of all the users and their collaborative levels for specific sensitive information.

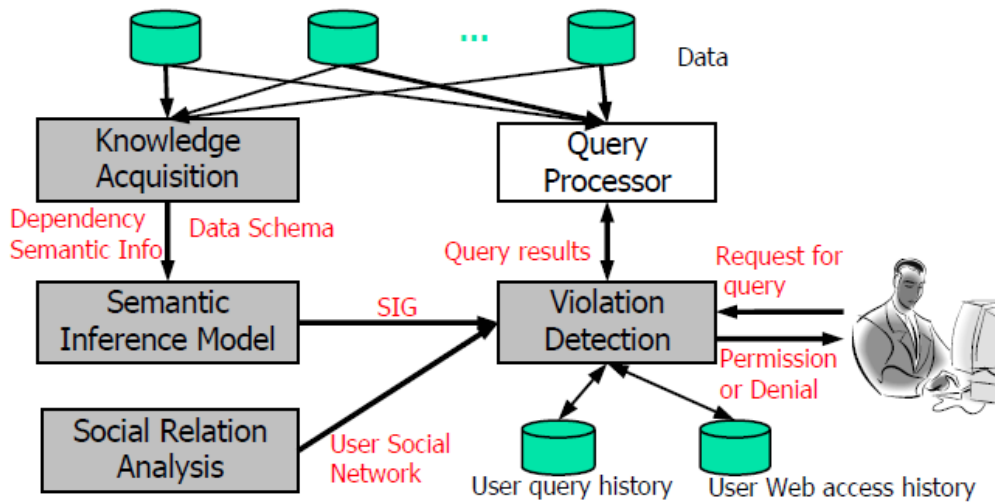


Figure 11: The framework for an inference detection system proposed in [32].

Furthermore, most of the work addressing inference detection for database systems are based on functional dependencies in the database schema. The authors of [34] show that in order to detect inferences more adequately, the data itself must also be considered. To make this happen, five inference rules were identified: *subsume*, *unique characteristic*, *overlapping*, *complementary*, and *functional dependency* inference rules. A prototype has been developed to detect if a user can indirectly access data using two or more queries. Improvement in the

system can be made by using distributed computing techniques to apply them on the queries in parallel or by employing any data access detection as an anomaly detection system.

In multilevel database, polyinstantiation occurs when different tuples with the same key, each at a different classification level, are allowed. However, there are problems limiting the use of polyinstantiation. To avoid these problems, a cover story is used for hiding the existence of a sensitive data. Therefore, a new technique is defined in [35] for managing cover stories free of semantic ambiguities, and offering a powerful security policy. The consistency and the security property of the multilevel database containing cover stories are also provided for the purpose of controlling database updates and deletions.

II.2 Ontology-based Inference

To formalize their privacy models, anonymization techniques [10, 13, 14, 36, 37] usually introduce adversary models which can conduct inference attacks in their systems. For instance, the works described in [38] and [39] use ontologies to preserve the individual's privacy in free text documents where data structure is missing.

In [38], the authors measure sensitivity of identifiable information through a top-down propagation technique using prefixed sensitivity levels mapped to a reference ontology. According to these computed sensitivity levels, words are disseminated. However, their approach considers only directly associated attributes (*quasi-identifying* entities, QIE) as keywords to a search engine. In essence, a common assumption in the privacy literature [40] is that there are three types of possibly overlapping sets of personal attributes:

- Identifying attributes: attributes, such as social security number (SSN), which identify a person uniquely,
- Quasi-identifying attributes: a combination of attributes which can identify a person uniquely, such as name and address,
- Sensitive attributes: attributes that users may like to keep hidden from the public, such as politic affiliation and sexual orientation.

In [39], the authors use a probabilistic algorithm to mine all searchable information concerning individuals. They use domain-specific ontologies to capture inferable information and eventually provide more accurate results. They focus on all searchable information about users on the Web. Therefore, they define three categories of personal information with different privacy-sensitive degree: *Identity information*, *privacy-Sensitive information*, and *other indirect information*.

Unfortunately, the ability of these techniques to deal with adversaries enforced with plausible background knowledge is limited when using domain-specific ontologies to compute sensitivity levels. These so-called levels of sensitivity should depend mainly on the knowledge that the adversary already has acquired which could be out of scope of a specific ontology.

II.3 Web-based Inference

As we mentioned earlier, the problem of inference control comes from the fact that the knowledge that can be extracted from the union of the private and reference collections is greater than the union of what can be extracted separately from the two collections. To cope with this, it becomes important to determine what information can be released publicly without compromising certain secrets, and what subset of the information cannot be released.

With the explosion of user-contributed content on the Web, using online (multimedia) content can be helpful to detect inferences. However, it is exposed to uncertainty since it is originated from many sources. To overcome this imperfection, Schoenmackers et al. present HOLMES: a knowledge-based model construction [41]. To construct HOLMES, they combine information from multiple web pages to infer assertions not explicitly seen in the textual corpus. In order to produce the probabilistic network, they write inference rules as Markov logic Horn clauses [42]. In [43], they proposed the SHERLOCK system to avoid labor and expertise in hand-crafting the appropriate set of Horn clauses for each new domain.

Word co-occurrence information is also important to extract personal information from the Web in different contexts. In [44], the authors propose a novel technique based on relevant occurrences to find user semantics (i.e., personal information) on the Web. Nevertheless, the amount of information is not always a relevant measure of dependency. For instance, two “tweets” with minimum co-occurrence might be issued by the same individual. Therefore, the authors in [45] suggest that what may be of critical importance are phrases that return only a *few* results from search engines. Their method is based on fingerprints of information: cyclical hashing is used to split the document to multiple parts. They use search engines (1) to check the popularity of phrases and (2) to detect inferences between keywords. Their hypothesis is that if the search engine returns only a few results for the search, the phrase is then considered sensitive. The authors in [33] consider both significant-frequent and significant-rare keywords.

Staddon et al. [31] initially based their work on an outside knowledge to detect clues of inference. They suggest a Web-based solution to control undesired inferences. They derive new knowledge by combining existing pieces of knowledge. They extract relevant keywords from the document to be published, and query the web in order to capture additional knowledge contributing to a privacy breach. Important improvements could include Web caching, besides an inference detection tool capable of functioning in real-time to significantly improve performance. In [46], the same authors refined their model of Web-based inference detection. They assume the existence of rules which specify how to derive new knowledge from the combination of existing pieces of knowledge. Their model finds all associations of a sensitive nature. They first search for the most frequent association of topic-based terms using search engines. Their findings are not limited to rules with large support. In essence, they also consider inference from low support patterns, since sensitive information has more likely a low support. In [47], they present an attribute-based encryption scheme for document redaction. They consider sensitive personal information, and also group of personal attributes that may indirectly allow inferring a person’s identity.

In [48], the authors present the notion of k -safety in which the identifying terms should be associated with at least k individuals. The authors in [49] sanitize sensitive parts of the

document to measure information loss and risk disclosure. They assume that a relevant sanitizing process could be applied to maintain the utility of information in the document.

Security requirements should also be achieved on XML documents since XML is adopted in data-publishing applications. Sensitive information could be leaked during XML publishing if common knowledge (constraints) is not considered carefully. Hence, the goal of the work in [50] is to protect sensitive information in the presence of data inference with common knowledge. This work formulates the process that users can infer data in XML publishing by using 3 types of common XML constraints: a child constraint, a descendant constraint, and a functional dependency. Towards that goal, one of the major challenges is finding a partial document of a given XML document without causing information leakage, while allowing publishing as much data as possible. It is also proved that there is a unique maximal document users can infer using the constraints, which contains all possible inferred documents.

II.4 Discussion

In Table 1, we summarize the different categories of inference control approaches with a description of the input and data types they use and the output they generate.

Table 1: Comparison of inference control approaches.

Approaches		Background knowledge	Data type	Output
Database inference	<i>k</i> -anonymity privacy protection using generalization and suppression [36]	External information from structured relational datasets	Text	Person-specific data that cannot be linked to other information
	<i>k</i> -anonymity: a model for protecting privacy [30]		Text	Private data where the subjects of the data cannot be re-identified
Ontology-based Inference	Privacy measures for free text documents [38]	Domain specific ontologies	Text	Score of sensitivity of information
	Preserving privacy on the searchable Internet [39]		Text	Inferable information about an individual
Web-based Inference	Efficient techniques for document sanitization [48]	Context of words / entities	Text	Minimum terms to be removed for the document to be sanitized
	Finding user semantics on the web using word co-occurrence information [44]		Text	User's semantics
	Web-based inference Detection [31]	Public web documents, Sensitive keywords	Text	Inferences that can be drawn about the user
	Large online social footprints [16]	User's pseudonym(s)	Text	User's information from social networks
	Studying user footprints in different online social networks [21]	User's profile URL	Text + Image	User's digital footprints

Several techniques have been defined in the literature [27-30] to prevent information disclosure and eliminate possible linking attacks that are used for individual re-identification.

These techniques assume that identifiable information and adversaries' background knowledge are stored in structured relational datasets. Specifically, they address linking attacks that can be established between (quasi)-identifying and sensitive attributes of individuals stored in schema-based relational tables without referring to multimedia objects such as images and videos. In addition, ontology-based techniques use only domain-specific ontologies to capture inferable information.

Web based techniques propose web-based solutions to control undesired inferences. However, their ability to handle multimedia documents is limited. The only few techniques [38, 39] proposed to handle identity anonymization in multimedia documents assume textual data with no reference whatsoever to multimedia objects such as images and videos. To summarize, two main challenges are involved when addressing linking attacks on social networks:

- 1- **Multimedia data:** Social networks have become a popular way to disseminate different types of information (e.g., text, video, audio, photo, etc.). To prevent information leakage, privacy-preserving algorithms should consider multimedia data and their associated metadata.
- 2- **Background knowledge:** Data shared on social networks (e.g., photos, videos, statuses, etc.) may include private and sensitive information. Such information may be helpful to the adversary to discover additional information that the user did not provide in his online profile for privacy reasons, and thus may be needed for inference detection.

In our work, we cover not only personal information provided as text directly by the user on his profile. Our approach includes other information that are most likely related to the user but in an indirect way through multimedia content in his own posts as well as those shared by his contacts online.

III. Online Identity

The concept of identity on the Web is not new, although terminologies vary [16, 20, 21, 51]. In the following, we first review some of the existing definitions of digital identity. Then, based on our classification, we describe data types used in each approach and show how they construct an individual's online identity and some of their limitations.

III.1 Online Identity Terminologies

The authors of [16] introduced the concept of *Online Social Footprint* (OSF). A user's online social footprint is the total amount of personal information that can be gathered about an online identity by aggregating his online social networks (OSN) profiles. To illustrate, Figure 12 shows a user named Bob Smith and his online social footprint, which was constructed with information from three social networking sites. The combination view of Bob's information from all these sites represents his online social footprint.

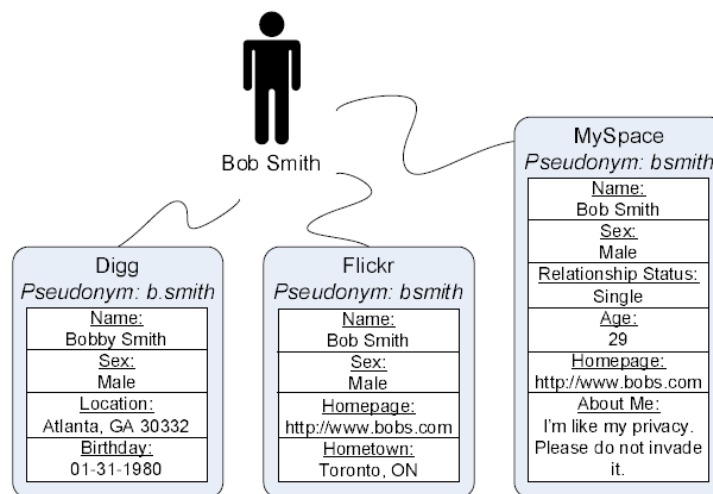


Figure 12: Online Social Footprint defined in [16].

The authors of [17] used the concept of OSF to quantify the amount of information revealed across multiple OSNs. This description was followed by a definition of a *Digital Footprint* [21] which represents the set of all personal information related to a user on online social

sites. Information can be either provided by the user directly or extracted by observing the user's interaction within several social sites.

Based on background knowledge from Web 2.0 platforms, the authors of [20] adopted a disambiguation approach to build a unary set of web resources containing personal information that refer to a given person and which they call *Identity Web References*.

The authors of the di.me project³² defined the *Personal Information Sphere* (PS) which refers to all the digital forms of information related to a user and that are available on the user's personal devices and online accounts (e.g., files, folder structures, contact lists, photo albums, status messages, etc.).

Cortis et al. [25] define a user's *online profile* as the set of personal data stored on distinguishable online accounts (networking platforms, email, instant messaging, calendaring, task management, etc.). Personal data in these accounts vary from static identity-related attributes (e.g., name, location, picture-url, phone-number, etc.) to more dynamic information such as online posts on social networks. It also includes information about the user's contacts in his social network: their identities and posts they shared are also retrieved. All these kinds of personal data from *one* of the user's accounts form his "online profile", whereas the aggregation of his personal data from *all* his accounts forms his "super profile".

After reviewing various definitions found in the literature, we classify existing methods of building online identity for into three categories:

- Statistical approaches,
- Learning-based approaches, and
- Ontological approaches.

³² di.me = "digital.me: Userware for the Intelligent, Intuitive, and Trust-Enhancing Management of the Users Personal Information Sphere in Digital and Social Environments" <http://www.dime-project.eu/>

III.2 Statistical Approaches

First, we present some statistical approaches that deal with the problem of online identity. In [16], the authors used online identity management sites to construct users' social footprints. Furthermore, the authors of [17] explored 40 types of attributes representing personal information available in public online profiles (e.g., *age*, *email*, *links*, *location*, *name*, *religion*, *user name*, etc.). The authors of [18] included images in their crawled profile attributes and tagging activity data. In the following, we will outline the methodology used in each of these approaches.

III.2.1 Online Social Footprint

First, online identity management sites are social networking sites where users can create unique profiles and provide links to all their social network profiles in order to manage their online identity (e.g., ClaimId³³, FindMeOn³⁴, FreeYourID³⁵, MyOpenID³⁶). The aim of the work in [16] is to show the threat of private information leakage by combining information about a user from multiple sites. Briefly, the authors aggregate personal information of a user from various social networks to construct his *online social footprint*. They assume the attacker has prior knowledge of one or more pseudonyms of the user. If not, the attacker attempts to infer it from the user's real name with a pseudonym guessing method. With this prior knowledge, they investigate the threat associated with an online social footprint by measuring (i) its size and (ii) the ability of an attacker to reconstruct it based on the user's pseudonym(s).

This study shed the light on the link between a user's pseudonym(s) and his personal information. The authors showed that an attack on targeted individuals is possible using their pseudonyms without having to use network based de-anonymization techniques [52, 53]. Pseudonyms have the potential to reconstruct a user's social footprint. They can also be linked

³³ <http://claimid.com/>

³⁴ <http://www.findmeon.com/relaunch/>

³⁵ <http://freeyourid.com/>

³⁶ <https://www.myopenid.com/>

to the profile attributes of the user. While multimedia content is emerging on social networks, their study has the limitation of excluding images published on social networks. The generation of online social footprints includes only text-based attributes (e.g., *birthday*, *homepage*, *hometown*, *location*, *name*, *sex*, etc.). Hence, the information leakage model does not measure information that can be discovered from images. The process of matching profiles also ignores different types of posts such as free text entry fields, location and images. In addition, the authors focus on inferring the pseudonym from the real name only. They don't consider the potential of inferring pseudonym using other personal attributes (e.g., *location*, *hometown*, *birth year*, etc.).

III.2.2 Increased Online Social Footprint

To further account for the threat to privacy in social networks, the authors of [17] investigate various factors that contribute to increased *online social footprint*, previously introduced in [16]. They explored 40 types of attributes representing personal information available in public online profiles (Figure 13).

about me/bio	email	links	political
age	ethnicity	lived	quote
anniversary	family	location	religion
birth date	gender	movies	schools
body type	home town	music	smoke/drink
books	income	name	sports
children	industry	occupation	status
companies	interests	organization	tv shows
connections	interested in	orientation	user name
education	language	phone	zodiac sign

Figure 13: Attribute list of personal information explored in [17].

The aim of this work [17] is to show how aggregation of information from multiple OSNs can contribute to privacy loss. Three analyses have been done to examine and measure the amount of information that can be obtained by combining multiple OSN profiles.

First, the authors measured the average number of publicly available attributes in each of the OSN individually and when aggregated across multiple OSNs. Their findings show that the average online social footprint size steadily increases with the number of OSN profiles. This is due to three aspects:

- The variety of distinct attributes among OSNs;
- The diversity of privacy practices applied by the same user to the same information on different OSNs;
- The correlation between the practice of having many online profiles and the tendency to share information online.

Second, the authors examined the online social footprints based on four user demographic information: i) gender, ii) country of residence, iii) age and iv) occupation. The authors did not find a significant correlation between footprint size and the first three attributes. Nevertheless, the present study shows that the users' age can influence the tendency to share specific attributes³⁷. The user's occupation can also influence the amount of sharing (e.g., an increased disclosure of personal information by users with customer-facing occupations).

Third, and similarly to [16], the authors of [17] investigated the footprint size based on the use of pseudonyms. To do this, they applied string matching and regular expression techniques with Jaro-Wrinkler³⁸ distance metric to identify pseudonyms. According to their observations, users who use their real names (namely users on Facebook, Google and Flickr) have a larger footprint than those who use pseudonyms (namely users on Youtube, Blogger³⁹ and LiveJournal⁴⁰).

³⁷ Older users are more likely to provide religion and political view on Facebook, whereas younger users are more interested in sharing their favorite music and books.

³⁸ http://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance

³⁹ www.blogger.com/

⁴⁰ <http://www.livejournal.com/>

These measures were computed on the profiles initially collected. Observations showed how aggregation of information across OSNs allows gathering additional information through:

- i. Complementary: more than half of the set of combined attributes from any two OSN profiles of the same user are complementary to each other.
- ii. Consistency of information revelation: users have preferred patterns when revealing information across different OSNs.
- iii. Consistency of attribute values: users' profiles have a significant matching between their attribute values at different levels of granularity.

The drawback of this method is that it focuses only on information entered by the user in his profile and does not capture personal information from his personal images. Exploring images in profiles would contribute to richer online social footprint, based on the metadata of the image – objects, associated text and data. Adding metadata of the image to the cross-OSN analysis allows gathering further information.

III.2.3 Digital Footprints

In [21], the authors generate the user's online digital footprints by aggregating personal information related to him across social networks. To do so, they propose an automated supervised classification approach for matching profiles belonging to the same user from the Twitter and LinkedIn profiles. Their system architecture is depicted in Figure 14.

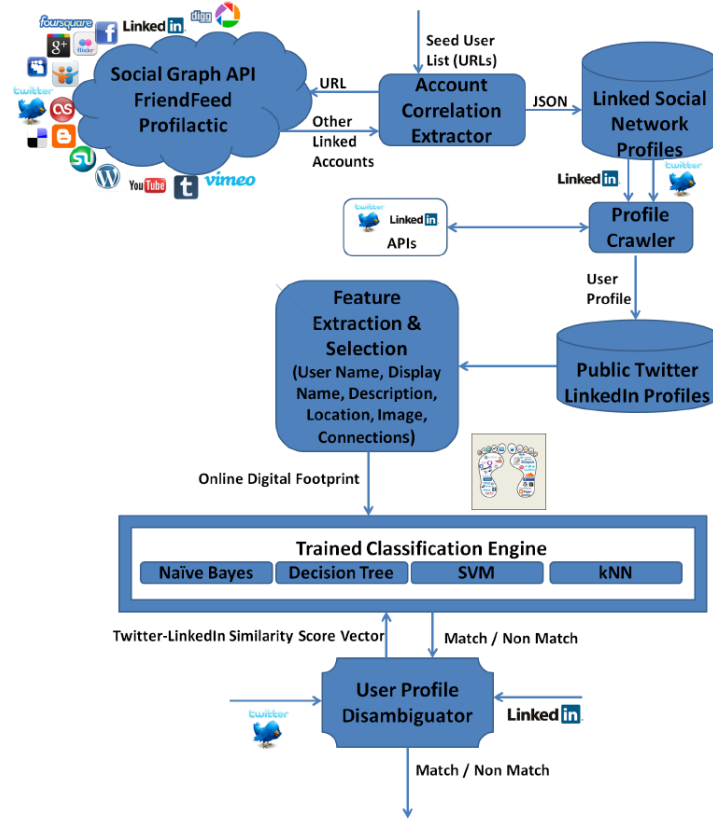


Figure 14: System architecture presented in [21].

Matching profiles is computed using feature-specific similarity measures: string, location and image matching. The problem of this approach is that their image analysis is limited to low-level visual features (i.e., color), thereby ignoring objects in the image (e.g., faces, places, objects of interest, text, etc.), metadata and textual information accompanying the image. Consequently the classification engine matches the pixel values between two images rather than information embedded in the image content and context (metadata files, description text, comments, tags, etc.). The matching criteria they use are restricted to statistical and engineering measures such as the Mean Square Error⁴¹, Peak Signal-to-Noise Ratio⁴² and the

⁴¹ http://en.wikipedia.org/wiki/Mean_squared_error

⁴² http://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio

Levenshtein distance⁴³. Thus, their classification does not rely on any semantic interpretation of the image.

In addition, they considered the number of connections of users for disambiguating the user profiles across social networks. This follows the intuition that a same user should have a similar number of friends in different social networks. This is generally not true due to the different nature of the services. However, examining online posts of the connections (friends) of a user is not considered in their study. Actually, information leakage is not strongly related to self-disclosure, as pointed in [54]. Friends may reveal a user's identifying information unintentionally no matter how much the user tries to protect his identity online. Thus, attackers do not just look at the user's profiles but also at friends of the user [55] (friends' profiles and posts, and more particularly friend's photos). Second, the limitation of the comparison method is that it does not involve cross-feature comparison: they compare values only between two same features from two profiles and do not associate values from different features across the profiles. For instance, the "location" feature extracted from a first profile can be matched with an element in the "description" feature in the second profile. Third, in their first analysis on data collected from Twitter, Youtube, Flickr and LinkedIn, they found that at least 80% of the profiles contain a profile picture. Results show much higher percentages (99%) of Twitter and Flickr users providing their profile images due to the nature of these services. The high percentages suggest that profile images would be needed for the digital identity construction, which is in line with the assumptions in our study.

III.3 Learning-based Approaches

Image features were used in several learning-based approaches [19-21]. For instance, in the *Digital Footprint* approach [21], a classification engine is trained with feature vectors including string attributes, location attributes and images. As for the authors of [20], they

⁴³ http://en.wikipedia.org/wiki/Levenshtein_distance

present a semi-supervised approach to automatically disambiguate the identity web references for a specified person. We will review their approach in what follows.

III.3.1 Identity Web References

A web resource, as defined by the authors, is any document on the web which is accessible by a URI such as web pages or data feeds. A web resource that contains personal information of a web user is an *identity web references* of the user. The aim of the approach proposed in [20] is to find all the web resources residing on the web which are *identity web references* of a given person. They rely on metadata descriptions from web resources (e.g., *name*, *email*, *website*, *location*, etc.) with no consideration of image metadata that can also contain personal information.

After retrieving the web resources that point to a particular person, it is important to know if they actually refer to that person. In order to classify the web resources as positive or negative sets, the authors of [20] proposed two disambiguation techniques:

- 1) a semi-supervised machine learning technique and
- 2) a graph-based technique.

The first disambiguation technique employs a self-training strategy. The choice of a self-training strategy is due to the limitation in the amount of the seed data. Instances from the initial seed data are first supplied to the machine learning classifier. Then a classification model is constructed based on the similarity of their features.

The second disambiguation technique applies a Random Walks [56] approach on the interlinked graph space obtained from the seed data and the set of web resources. Using distance measures on the graph, this technique clusters the web resources into positive and negative in order to find the identity web references. The overview of their approach is depicted in Figure 15: Overview of the approach proposed in [20] to disambiguate identity web references using Web 2.0 data..

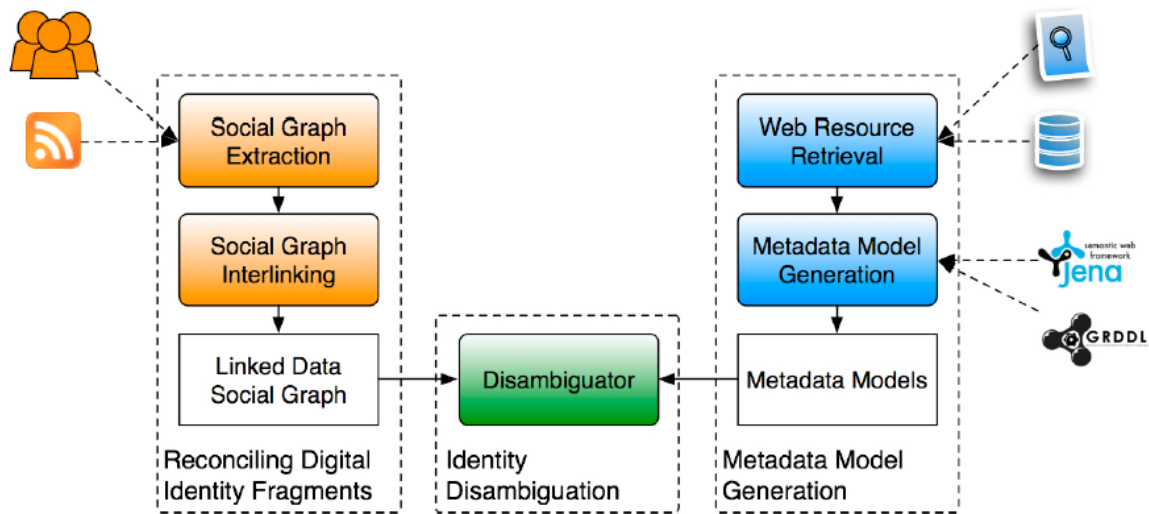


Figure 15: Overview of the approach proposed in [20] to disambiguate identity web references using Web 2.0 data.

In this approach, disambiguation techniques achieve consistent accuracy with all web presence levels. Yet, regarding content types (images, videos, audios) of the web presence is essential given the trend of publishing and sharing multimedia information on social web sites and online communities. In addition, the background knowledge used for learning is generated by exporting and combining data from multiple heterogeneous online accounts such as Facebook, LinkedIn and Twitter. However, the use of these Web 2.0 platforms in this approach does not cover the maximum knowledge about a web presence. Indeed, metadata extracted from online images of a given person is able to reproduce larger presence of the person on the web. The disambiguation approach must be able to cope with multimedia resources, such as images as seed data to start the identification process.

III.3.2 Enhanced Identity Merging based on Face Recognition

Another approach [19] using images shows that profile images can provide additional power in the user re-identification. In [19], the authors combine traditional text-based and image-based approaches to enhance available user re-identification systems for social network data aggregation. Their method uses a face-recognition algorithm to re-identify user profiles based on their profile images. The advantage of this method is that it incorporates images with traditional text-based. Nevertheless, this method uses face-recognition software to compare the images uploaded by users without considering image metadata. An overview of their method is illustrated in Figure 16.

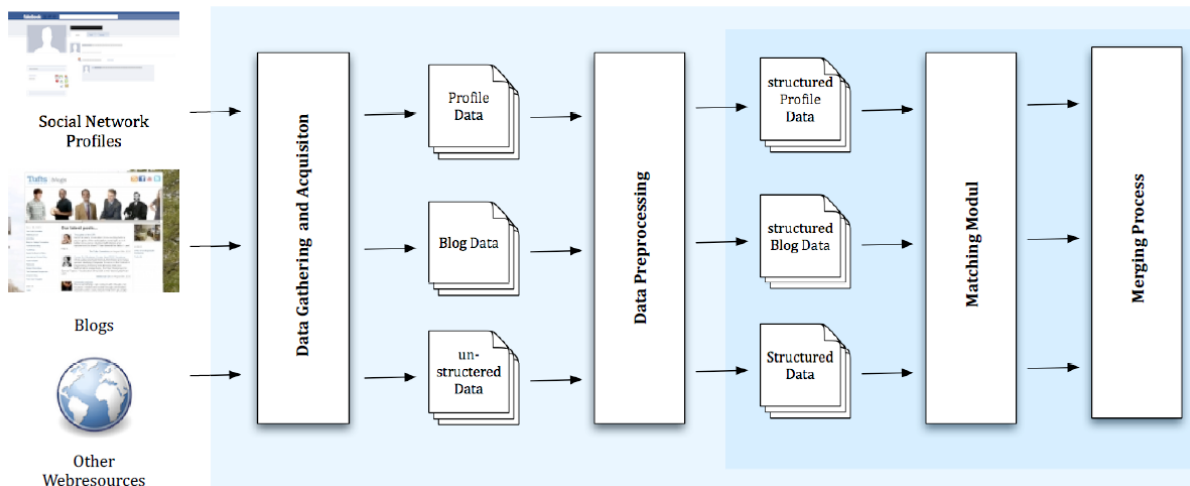


Figure 16: Approach used in [19] for data aggregation.

III.4 Ontological Approaches

Different works [22-26] used ontologies to attribute semantics to the online identity. We start by providing a brief overview of ontologies designed for managing user presence on the Semantic Web, then we detail the most relevant ones.

Some ontologies, initially designed for desktop personal data in the NEPOMUK project [57], are useful for personal data available on the Semantic Web. Examples of these ontologies are: the NEPOMUK contact ontology⁴⁴ (NCO), the Information Element ontology⁴⁵ (NIE), the Personal Information Model Ontology⁴⁶ (PIMO), the Account Ontology⁴⁷ (DAO) and the Annotation Ontology⁴⁸ (NAO).

Other ontologies were developed for describing personal profile. The FOAF vocabulary⁴⁹ represents individuals and relationships between them. It defines the typical attributes that describe the individuals' identity, their online coordinates and basic web activities [58]. Then, the Semantically-Interlinked Online Communities (SIOC) ontology [59] reused the FOAF vocabulary for exploring personal profile attributes and social networking information (the content the user produces and the actions of other users on this content).

A mapping approach between digital identity ontologies, named the *Social Identity Schema Mapping* (SISM) is proposed in [26] to address the problem of information heterogeneity.

The *LivePost Ontology* (DLPO) is designed in [23] to represent online posts shared across social networks and their embedded knowledge. More importantly, personal information in DLPO is gathered from different types of posts within different social networks: *Message*, *MultimediaPost*, *WebDocumentPost* and *PresencePost*. DLPO integrates concepts from PIM models and existing standards namely the SIOC ontology and several domain ontologies (i.e., NIE, PIMO, DAO, and NAO).

The most relevant ontologies used are detailed below.

⁴⁴ <http://www.semanticdesktop.org/ontologies/nco/http://www.semanticdesktop.org/ontologies/nco/>

⁴⁵ <http://www.semanticdesktop.org/ontologies/nie/>

⁴⁶ <http://www.semanticdesktop.org/ontologies/2007/11/01/pimo/>

⁴⁷ <http://sourceforge.net/apps/trac/oscaf/ticket/129>

⁴⁸ <http://www.semanticdesktop.org/ontologies/nao/>

⁴⁹ <http://www.foaf-project.org/>

III.4.1 The Personal Information Model (PIMO)

The Personal Information Model of a user (PIMO) [24] was proposed in the NEPOMUK project [57] and is now maintained by OSCAF members as a response to the need of having a formal representation of the users Mental Model⁵⁰. The EPOS project uses the PIMO to create the user's own domain ontology, called PIMO-User [60]. Personal concepts, ideas, projects, contacts and other resources that are of direct interest to the user represent the user's Personal Information Models (PIM). The PIMO⁵¹ ontology is simultaneously an RDF vocabulary to express these models and an upper ontology defining basic classes and properties to use. This ontology satisfies the following requirements: (1) precise representation, (2) extensibility, (3) interoperability, (4) reuse of existing ontologies, and (5) data integration [61]. The focus in PIMO is on data that is accessed through a Semantic Desktop or other personalized Semantic Web applications.

III.4.2 The Nepomuk Contact Ontology (NCO)

With the OSCAF/NEPOMUK ontologies, the user can cope with different formats of data. For instance, contact information is described in the NEPOMUK contact ontology⁵² (NCO). The meaning of the term "Contact" in NCO is quite wide. It is every piece of data that identifies an entity or provides means to communicate with it.

III.4.3 FOAF

FOAF is an ontology⁵³ defined in the FOAF project⁵⁴ to represent individuals and relationships between them. It defines the typical attributes of an agent that describe his identity, online coordinates and basic web-based activities. The FOAF ontology is a potential

⁵⁰ PIMO is based on the idea that users have a mental model to categorize their environment. It represents the user himself and the fact that he has a Personal Information Model.

⁵¹ http://www.semanticdesktop.org/ontologies/2007/11/01/pimo/v1.1/pimo_v1.1.pdf

⁵² <http://www.semanticdesktop.org/ontologies/2007/11/01/pimo/#>

⁵³ <http://www.semanticdesktop.org/ontologies/nco/>

⁵⁴ D. Brickley and L. Miller. FOAF: the 'friend of a friend' vocabulary. <http://xmlns.com/foaf/0.1/>, 2004.

⁵⁴ "The friend of a friend (foaf) project," <http://www.foaf-project.org/>.

tool for the extraction and the fusion of personal information: on one hand, it enriches the expression of personal information and relationships represented in RDF/XML as pointed in [20], including name, mailbox, homepage URL, friends, and so on. On the other hand, it provides a standardized format [62] using the Ontology Web Language (OWL), thus providing the capability to aggregate information from different sources.

III.4.4 Social Identity Schema Mapping (SISM)

The Social Identity Schema Mapping (SISM) [26] is a mapping vocabulary between five digital identity ontologies: The Friend of a Friend ontology⁵⁵ (FOAF), the ontology for VCards⁵⁶, the XFN ontology⁵⁷, the Personal Information Model Ontology⁵⁸ (PIMO) and the Nepomuk Contact Ontology⁵⁹ (NCO). It aims at mapping concepts from these ontologies in an interpretable and machine-readable format. The basis mapping in SISM is a subject-predicate-object triple: a (parsed) source concept has a relation with a (known) target concept, where the relation describes the semantics of the mapping. To fully capture such semantics, they combine mapping constructs from the OWL⁶⁰ and the SKOS⁶¹ languages. While the OWL language allows to describe the semantics of knowledge with concepts and relationships between concepts, the SKOS ontology is a vocabulary for organizing concepts with lightweight semantic properties (e.g., *skos:narrower*, *skos:broader*, *skos:related*). Relations⁶² defined in the SISM ontology enable to interpret in a common and precise way, metadata models containing identity information on the Web. SISM supports inference rules to normalize metadata models. Consequently, the SISM vocabulary enables the interlinking of identity fragments distributed on different social web platforms. *SISM* is used in [20] for normalization of metadata from different social web platforms for producing the *Identity Web References*.

⁵⁵ <http://xmlns.com/foaf/spec/>

⁵⁶ <http://www.w3.org/2006/vcard/ns>

⁵⁷ <http://vocab.sindice.com/xfn>

⁵⁸ <http://www.semanticdesktop.org/ontologies/2007/11/01/pimo/#>

⁵⁹ <http://www.semanticdesktop.org/ontologies/2007/03/22/nco/>

⁶⁰ Smith, M., Welty, C., and McGuinness, D.L.: OWL Web Ontology Guide. W3C Recommendation. <http://www.w3.org/TR/owl-guide/> (2004)

⁶¹ Isaac, A. and Summers, E.: SKOS Simple Knowledge Organization System Primer. W3C Recommendation. <http://www.w3.org/TR/skos-primer/> (2009)

⁶² Relations are: equivalence, hierarchical and associative

III.4.5 The LivePost Ontology (DLPO)

The LivePost Ontology (DLPO) [23] is an open knowledge representation standard which models online posts shared across social networks. The main purpose of the DLPO is to capture implicit presence knowledge embedded in the social web sharing activities (e.g., online posts, online interactions, and online Social Web practices, etc.). This ontology is part of the *di.me* project [63] which aims at unifying the user's personal information across various heterogeneous sources (Figure 17). Personal information is gathered from different types of posts within different social networks. It consists of four types of posts: *Message*, *MultimediaPost*, *WebDocumentPost* and *PresencePost*. A post item can occur individually or in conjunction with other post items. The DLPO integrates concepts from PIM models and existing standards namely the Semantically-Interlinked Online Communities (SIOC) ontology [22] and many domain ontologies (e.g., the Information Element ontology⁶³ (NIE), the Personal Information Model Ontology⁶⁴ (PIMO), the Account Ontology⁶⁵ (DAO), the Annotation Ontology⁶⁶ (NAO)). More specifically, the SIOC ontology defines primitives for describing a user in social web sites and online communities, the content he produces and the actions of other users on this content. NIE, PIMO, DAO and NAO are part of the NEPOMUK⁶⁷ ontologies. Through the integration of these ontologies, the DLPO ontology enables to describe dynamic personal information in a concise and semantic way. It defines semantic links between:

- i. user posts from different social networks;
- ii. user posts and user presence⁶⁸;
- iii. user posts and global semantic data clouds⁶⁹.

⁶³ <http://www.semanticdesktop.org/ontologies/nie/>

⁶⁴ <http://www.semanticdesktop.org/ontologies/2007/11/01/pimo/>

⁶⁵ <http://sourceforge.net/apps/trac/oscaf/ticket/129>

⁶⁶ <http://www.semanticdesktop.org/ontologies/nao/>

⁶⁷ <http://www.semanticdesktop.org/ontologies/>

⁶⁸ The term "presence" refers to both online (activities e.g. check-ins, posts, liking; interactions e.g. playing a game, chatting; availability, visibility, etc.) and physical (activities e.g. travelling, walking, working; current location, nearby people and places, etc.) user experiences.

⁶⁹ <http://richard.cyganiak.de/2007/10/lod/>

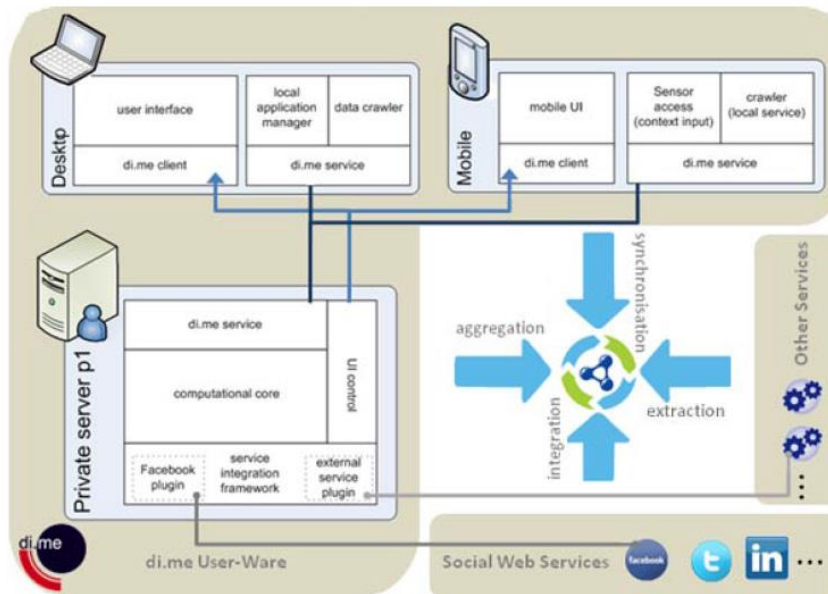


Figure 17: Overview of the distributed di.me system architecture in [63].

Table 2 shows a comparison of some digital identity ontologies with respect to the multimedia dimension and the data covered. The hand symbol in this table (and the others) means that the studied approach takes into consideration the given criteria.

Table 2: Comparison of ontological approaches.

Ontology name	Personal Information	User-generated content	Images	Metadata
PIMO [61]	✋	✋		
FOAF [58]	✋		✋	✋
SIOC [59]	✋	✋		
SISM [26]	✋			
DLPO [23]	✋	✋	✋	

III.4.6 Integration of multiple ontologies

The authors in [25] focus on the integration of user profiles from multiple heterogeneous online accounts such as Facebook, LinkedIn and Twitter. Since it is possible to have multiple occurrences of the same data across different accounts, providing an abstraction for data representation is needed to eliminate duplicate instances. To do so, they propose an ontology-based approach to detect semantically-equivalent user profiles, illustrated in Figure 18. User profiles are mapped into one ontology framework consisting of a set of re-used, extended as well as new vocabularies provided by the OSCA Foundation⁷⁰ (OSCAF). More specifically, they used the Personal Information Model (PIM), provided by PIMO, as the knowledge base of their approach. The PIM contains data that is of direct interest to the user. Initially, it is automatically populated by crawling data on the user's online accounts or devices. Further, the user can extend it manually by adding the representations of his own mental model. Since it is the user's own PIM, such knowledge base contains the most valuable information about the user. It is certainly smaller, leading to more accurate results.

⁷⁰ <http://www.oscaf.org/>. The "Open Semantic Collaboration Architecture Foundation" (OSCAF) is an industry led group bringing together organizations and individuals interested in ensuring interoperability between next generation desktops and collaborative environments.

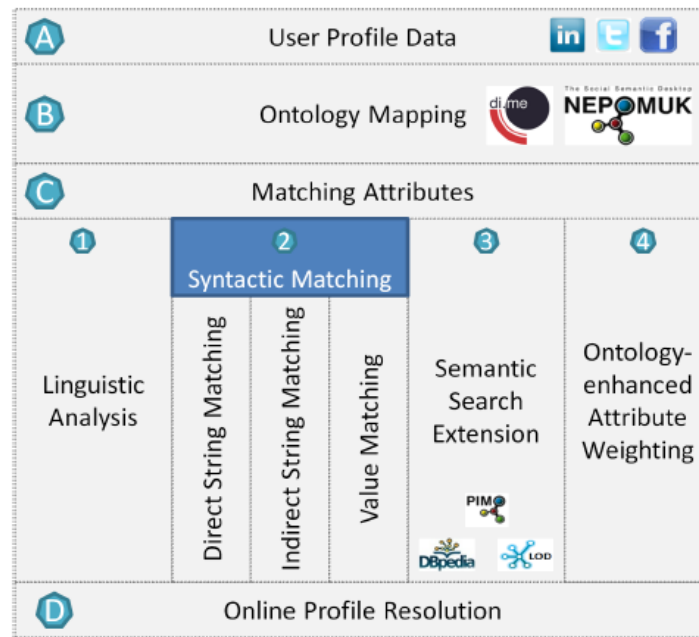


Figure 18: The approach process proposed in [25].

To conclude, this work [25] proposes an automated approach for the integration of several identities that are distributed across multiple local and remote sources. This approach facilitates the management of the multiple identities of a user with minimal effort. However, it has three limitations:

The NCO and the DLPO ontologies are able to express information extracted from images. First, in the NCO ontology, a photo (*nco:Image*) can be attached to the class *Contact*. Second, in the LivePost ontology, an image post (*dlpo:ImagePost*) can generate an instance of the *MultimediaPost* class.

In addition, the authors consider semantic relatedness only among text entities in the profile attributes to compute the user profile matching. The matching between the user's profile data and the user's PIM is limited to text (e.g. *surname*, *username*, *address*, *country*, etc.).

Although the NCO and the DLPO ontologies handle image data, this work [25] retrieves only textual attributes from a user profile, excluding multimedia content available online. The matching should also be extended to include images and consider semantic information from

their content (e.g., faces). Images can be either available in the online profile of the user or his contacts.























Furthermore, metadata matching targets only resources from a user's profile but lacks the exploitation of metadata that is attached to shared items (e.g., location coordinates from a posted image). We believe that adding the Nepomuk Exif Ontology (NEXIF) to the ontology framework would be useful for capturing information from the EXIF metadata of the user's images. Furthermore, NEXIF⁷¹ allows the EXIF metadata to be expressed in RDF.

III.5 Discussion

So far, we presented existing approaches regarding online identity construction, and described both their advantages and limitations. In Table 3, we summarize these approaches with a description of their features and limitations.

⁷¹ <http://test.semanticdesktop.org/ontologies/nexif.html>

Table 3: Comparison of online identity approaches.

	Online Social Footprint [16]	Online Social Footprint [17]	Digital Footprint [21]	Identity Web References [20]	Personal Information Sphere [25]	Tag-based Profile [18]
Text						
Image						
Metadata						
Entities/Concepts						
Persons						
Locations						
Time						
Association Text-Image						
Semantic mapping						

Existing approaches on digital identity provide solutions for the aggregation of personal information across multiple Web 2.0 platforms. Nevertheless, most studies are restricted to textual data. In the meantime, approaches including images are based on low-level features. Some methods [19] use key facial features to incorporate the visual content of images. However, they do not extract users' semantics from metadata attributes behind images (e.g., location, time, etc.). Most of these approaches ignore the potential of obtaining information from events.

Some approaches [22-26] used several ontologies for semantically describing the digital identity. Considering the complex and unstructured information, the use of ontologies is

important to explicit the concepts and their relationships between the user data. Ontologies have to include various content types of a web presence such as image representations and metadata.

In conclusion, while existing approaches investigate the ability of obtaining personal information using Web 2.0 data, two key differences distinguish our work:

- We use image metadata to build richer user profile information from social networks.
- We detect personal events in photos shared by the user or his connections in a social network in order to discover additional information that might not be included in the user profile.

IV. Events in Online Social Networks

Photos taken during events might contain references to personal data that the user may have not intended to disclose online. In this section, we review previous definitions in the literature, and then we present the problem of detecting events from photos shared in online social networks.

IV.1 Event Definitions

The first approaches of event detection proposed in the literature were associated with the Topic Detection and Tracking (TDT) [64]. The objective of TDT is to detect topically related stories within a stream of news media [65]. Basically, an event identifies something that happened in a certain place at a certain time [66]. This definition was adopted within the TDT project, then was developed to detect events from news stories coming from social media [67, 68]. For instance, *Knowle* is a news event centrality data management system for organizing news events on the Web [69]. A news event in the *Knowle* system is defined by a life course and a set of features describing the event.

With the emergency of multimedia applications, the concept of event is becoming one of the most challenging objectives in the semantic Web, especially the social Web.

Many different definitions of events have been proposed to process multimedia data [70], most of which are based on the spatial [71-73], temporal [69, 74] or spatio-temporal aspects [75-77]. Currently, there are several ways of defining an event in the literature.

Rattenburt et al. [74] regard an event as a specific segment of time. Paniagua et al. [78] represent an event as a pair of temporal and spatial information. The definitions by Mahata et al. [75] and Valkanas et al. [77] claim that an event is a real-world incident or phenomenon, that occurred at some specific time (or over a certain period of time), and is usually tied to a location. Sakaki et al. [79] refer to event as an arbitrary classification of a space/time region. It might have actively participating agents, passive factors, products and a location in space/time. Gupta et al. [80] identify an event using a set of required core words and a set of optional subordinate words. In [81], an event is defined as a 3-tuple $\langle E, R, t \rangle$, where E is a set of entities, $R \subseteq E \times E$ is a set of dynamic relationships, and t is a continuous time window. This definition enables discovering and verifying dynamic connections among entities that are connected at a given time period. These dynamic relationships are then used to derive dynamic events by identifying and enriching them with further information.

Other definitions have provided categories for events. For example, events can be classified as *local* when they refer to personal experiences (e.g., wedding, birthday celebrations, etc.), or *global* events (e.g., sport competitions, concert, natural disasters, etc.) [82]. *Global events* allow building collective experiences for sharing personal experiences as part of a social phenomenon called *collective events* [83]. Other categories are also provided in [82, 83] such as *home* and *away-from-home* events, *routine* and *non-routine* events in order to recognize the basic nature of an event. Figure 19 shows a distinction of personal and social event with regard to the memory and experience.

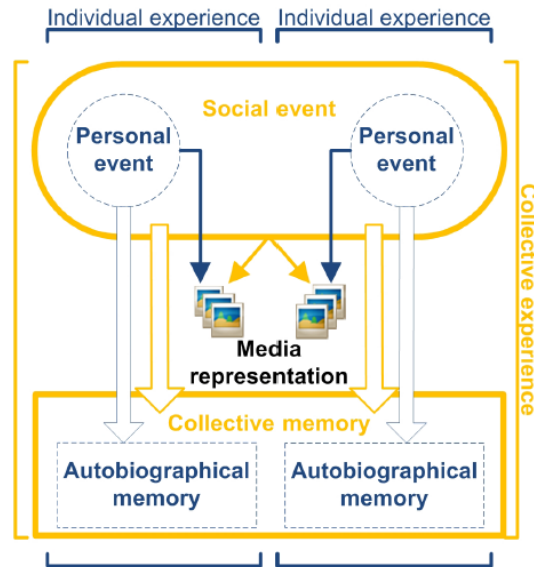


Figure 19: Distinction between personal events and social events, as defined in [83].

As a conclusion, there is no common definition of events established for multimedia applications. The multimedia event model presented in [84] shows the need of having a common multimedia event model for a wide diversity of applications, event-centric multimedia, reusable event management infrastructure and tools, and application integration. A common event model should be able to identify six distinct aspects of event description: temporal, spatial, informational, causal, structural, and experiential. Figure 20 illustrates these aspects as well as the determinants of each aspect. Also, many relationships can be expressed by using a suitable event model such as events' temporal and spatial relationships as well as structural and causal relationships.

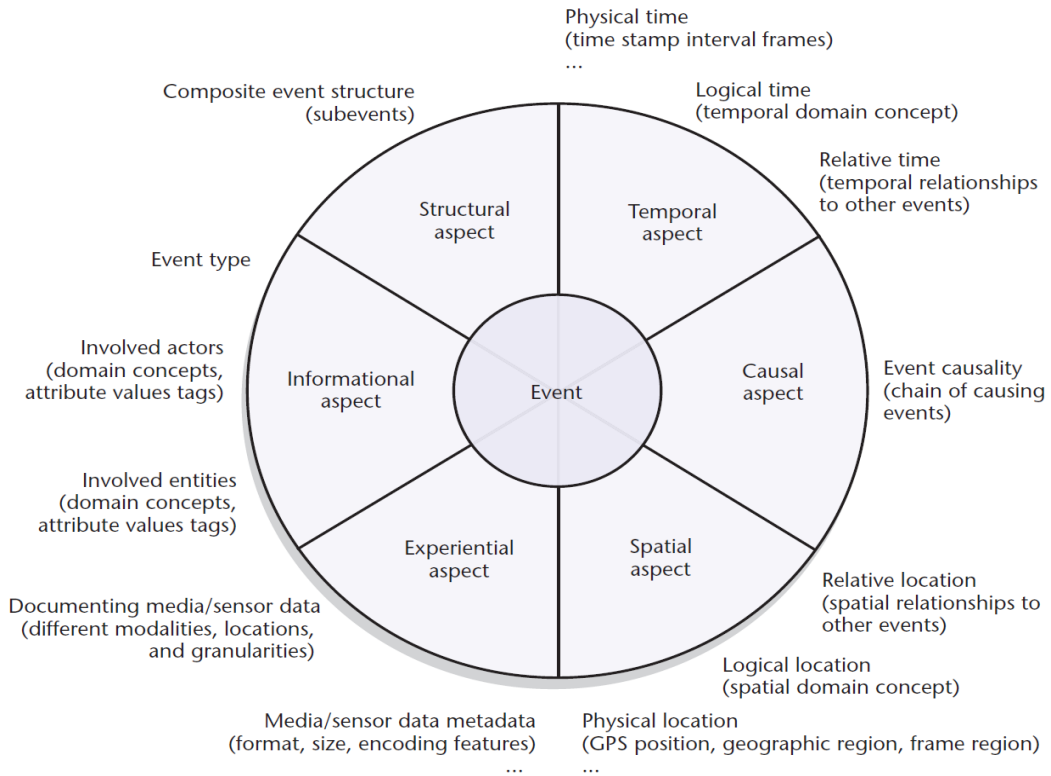


Figure 20: Basic aspects of event description from [84].

IV.2 Event Detection from Photos in Online Social Networks

A large number of studies investigated the detection of events from multimedia data including text [85-90], image [73, 78, 82, 83, 91], and video [92-95]. In our study, we are interested in approaches exploring image data. In essence, extracting events from multimedia in terms of photographs or images is much more difficult when compared to text. This is essentially due to two reasons, according to the authors of [96, 97]. First, event detection from images requires aggregation of heterogeneous metadata. Second, linking multimedia data to event model aspects is far more challenging than textual data. We focus here on methods for the detection of events from images taken from the real-life of an individual and shared on online social networks (i.e., personal events). In the following section, we discuss the main approaches and how they differ from our approach. We classify them in two groups: i)

clustering approaches [83, 91, 98-102] and ii) non-clustering approaches [69, 71, 72, 82, 103, 104].

IV.2.1 Clustering Approaches

New approaches have emerged in the past few years in the area of event detection from images on online social networks. For instance, the authors of [82, 83] use the context allied with social media content for the event detection.

The methodology in [82, 83] uses the context of social media content, user provided tags and significant terms for each event from the Internet as features for event detection. Their method, illustrated in Figure 21, groups the images to events simultaneously.

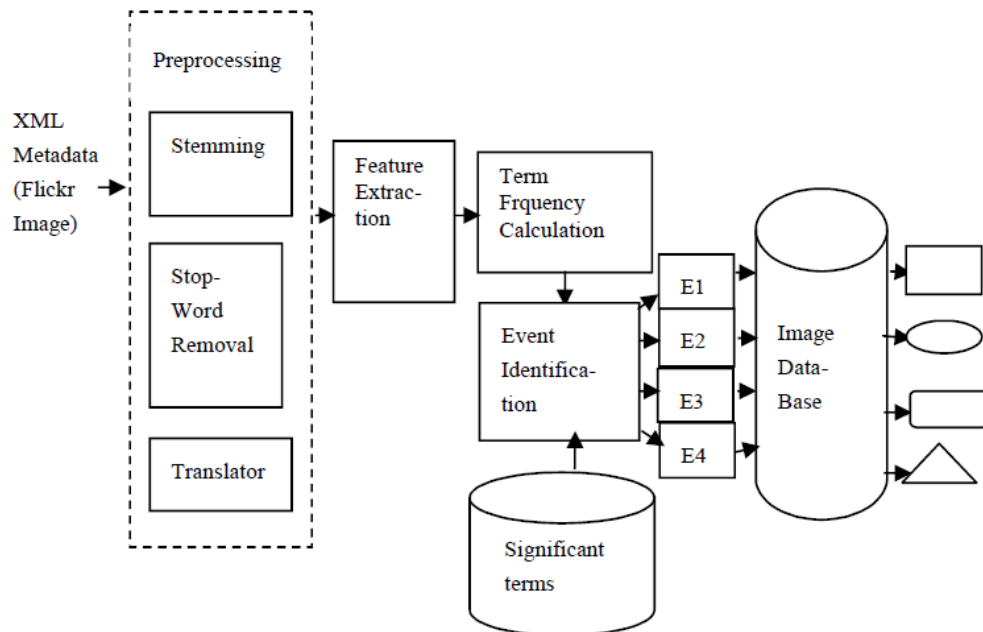


Figure 21: Event detection framework presented in [82, 83].

The authors of [83] refer to context as any information that can be captured by a camera, or extracted from related images in the same photo collection or any textual description provided by the user. The aim of their work is to detect and arrange events in private photo archives by building a contextual meaningful hierarchy of events. Therefore, they suggest an automatic image clustering that merges visual data with context information (time, space). In a subsequent work [78], the same authors seek to group events coming from different users into richer social events based on two metadata features (time, space). More specifically, they show how event detection can bring together users who participated in the same event and hence propagate social connections among users.

Similarly, the authors of [91] investigate photos posted on social media sites to detect social events. They use four types of heterogeneous metadata related to photos: time-stamp, location, visual data and textual description. First, they represent the social media and their metadata into a star-structured K -partite graph. Second, they model relationships between social media and relationships between metadata sets. Then, they apply a co-clustering method on the star structure.

In the work presented in [105], the authors propose a general and scalable online clustering framework for identifying events and their corresponding social media documents. In essence, social media sites represent a valuable source of event information, but this information is far from uniform in quality and might often be misleading or ambiguous. To address this problem, they use contextual data of social media documents and define similarities tailored to the social media domain. The similarity metric is the cosine similarity metric for textual features (e.g., title, description, etc.), and the Haversine distance for location metadata. Several existing techniques are applied in this work to learn multi-feature similarity metrics for social media documents. Since individual features might be noisy or unreliable, they suggest that all features, considered collectively, can provide more reliable information about events. Therefore, to improve the efficiency of the document clustering, they use first ensemble-based and then classification-based techniques to learn a combined similarity metric.

IV.2.2 Non-clustering Approaches

The authors of [82] refer to context as the *dateTaken*, *title*, *description*, *tags* (user comments), and *latitude* and *longitude* information. In their approach, they propose a method for grouping together photographs of similar events using significant terms for each event from Wikipedia⁷² and Google⁷³. However, they do not consider geolocation in their analysis because the geographic information is not always available in real-world dataset.

The authors of [73] detect events by temporally monitoring the social media sharing activity at specific locations. They use metadata attached to photos (tags, description, and geographic location) to enrich the event dataset and infer the topic of events. This work shows similarities with our work because of the metadata interest, but they do not detect links between events.

















IV.2.3 Discussion

In Table 4, we compare the above event detection approaches.

⁷² https://en.wikipedia.org/wiki/Main_Page

⁷³ <https://www.google.com/>

Table 4: A summary of photo-based event detection approaches.

Approach	Domain	Metadata	Text (description, tags)	Visual content	Other's resources	Event links	Dataset of experiments
[82]	Social media content						Flickr
[91]	News events						www.reuters.com
[83]	Personal photo collections						Private photo collection
[78]	Social networks discovery						Event-Media
[73]	Social media sharing						MediaEval SED

From the above table, one can see the following limitations:

- Insufficiency in exploring all image metadata: Most of these approaches do not include the creation metadata (e.g., creator name, geographic information, etc.). Some approaches consider visual data [73, 83, 91] by extracting only low-level features (e.g., color). Our model is able to make use of objects found in images, particularly faces of persons, since we assume that photos are annotated with person-tags.
- Deficiency of the multi-* property of events: Existing approaches do not apply to particular cases such as events that occurred many days, or events that took place in many cities, regions or countries. In our study, we address these particular cases by using the multi-* property of events, namely *multi-day*, *multi-site*, *multi-source* and *multi-participant* events.

- Restriction to the user's personal profile: Most of the existing approaches [73, 82, 83, 91] limit the extraction of images to the concerned user's profile only. In our approach, we exploit personal events in the individuals' own posts as well as those shared by their friends/contacts.
- Lack of event links: Existing approaches do not consider the different types of links between events that we propose in our work (temporal, spatial, social, and semantic). In the following, we review different approaches that have been developed to identify relations between events.

IV.3 Events Linking Approaches

In this section, we focus on event linking approaches [81, 84, 106-109] that allow to identify links between events. We summarize and discuss their limitations and features. Then, we try to highlight the difference between our approach and the existing ones.

The challenges that are faced when modeling real-world events captured in multimedia documents as discussed in [109] are the following:

- i) Event semantics are dynamic and might change over time, across people, and depending on the context.
- ii) People may upload and tag a multimedia document without associating it to events, which creates the problem of resolving the semantics of the document.
- iii) The relationship between the multimedia document and the event is driven by the user's need and hence may be defined by people's own words, languages and expressions.

In order to address these problems, the authors of [109] introduced *Eventory*, a media archive of events corresponding to activities and events of an academic community. *Eventory* allows the following operations: i) multimedia document upload, ii) event creation and iii) event

notification. The media upload supports the sharing of all types of media files including videos, photos as well as text documents. The event creation follows the standard facets of *who*, *where*, *when* and *what*. In *Eventory*, relationships can exist between events, between media, and between media and events. Since in this work [109], the authors refer to an event as a real-world occurrence that unfolds over space and time, the authors focus on temporal (e.g., *daily*, *weekly*, *monthly* and *yearly*) and spatial relations (e.g., *beside*). Beyond that, they allow the user to define his own semantic relationships arbitrarily. Thus, the semantics will be meaningful to the user. Constraints are also imposed by the system on the semantics of the relationship's specification.

In [81], the authors distinguished between traditional and dynamic relationships extracted from unstructured or semi-structured data sources. Traditional relationships are static predefined relationships (e.g., spousal relationship) that can be extracted based on language patterns. In contrast, dynamic relationships are not predefined by an existing schema. They are temporally defined (e.g., co-bursting) and are often a product of underlying real world events (e.g., relationships associated with news events). In this work [81], the authors focus on *dynamic relationships* that are formed due to real-world events. However, the difference with our method is that they generate relationships before deriving events. Figure 22 shows an architecture diagram of their approach.

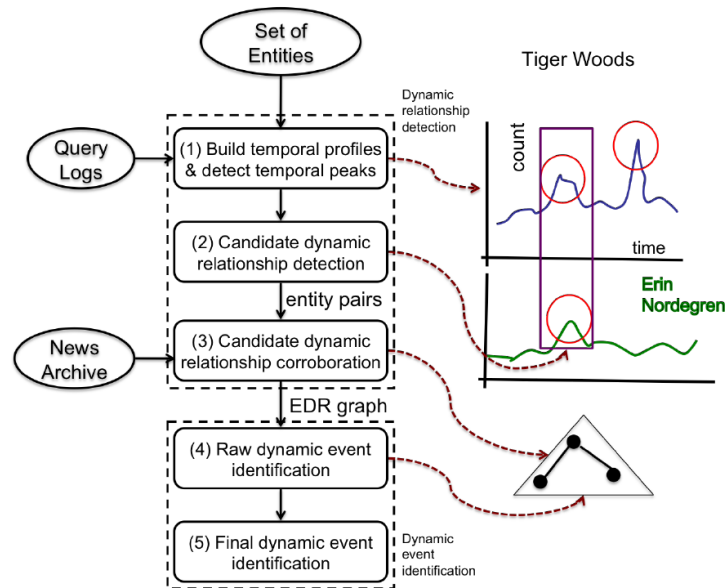


Figure 22: Architecture diagram of event detection approach proposed in [81].

Events are entity-based (an entity refers to any real world concept such as a person, a company, etc.), and concerned with dynamic relationships. This leads to the *entity dynamic relation graph*. Their approach starts by building the temporal profile of each entity by evaluating the number of documents that mention this entity within a specific time window. Candidate dynamic relationships are then detected using temporal profiling, co-occurrence and peak detection techniques [110]. The techniques employed apply to any granularity of event consolidation, though their experiments are based on weekly time windows. In the next step, they develop event identification algorithms to extract the dynamic events based on entity clusters with two temporal constraints (*global* and *local*) that they defined. Finally, they enrich events with three additional information: 1. *The entity involvement score* that measures the overall entity participation in the event; 2. *The event confidence score* that measures the probability of connectedness of two entities; and 3. *Descriptive information* about the events, including *event textual description* and *event popularity*⁷⁴.

⁷⁴ *Event confidence* refers to the confidence their system has of having correctly produced an event, while, *event popularity* refers to ranking events based on external importance assuming the event is correct.

In [106], the authors focus on the representation of video events and present a framework for semantic annotation of video, called EDF. The goal of EDF is to capture event semantics that enable storage, inference and retrieval of events from lower level event observations. The authors defined two classes of events: i) primitive event (comprised of a set of actors and actions), and ii) composite event (comprised of other primitive events). This framework allows a hierarchical decomposition of complex events into simpler ones. Based on top-level ontology, it supports reasoning and inference of composite events and relationships from the primitive ones. The ontology presented here consists of a set of predicates for describing spatio-temporal relationships between events and entities. The authors referred to *Allen's Interval Algebra* [76] to express temporal relations between two events. The spatial relationships in EDF are based on the typical spatial relations between any two objects: topological relationships [111], directional relationships [112], and distances (constraints on spatial metrics such as *distance* < 100).

EDF^E [113] is an extended version of EDF developed to facilitate the event detection process and reasoning about events. The extended version provides two additional features: i) the temporal predicate granularity as a feature of temporal association between sub-events, and ii) the event evidence to capture the full evidence for the detected events. EDF^E was applied to model complex events from real world surveillance videos in a retail store.

The authors of [108] focus on composite events based on the Event Model, called *E-model* [84]. However, composition operators are here restricted to only two facets: spatial and temporal. The authors subsequently defined a set of temporal and spatial operators independent of domains to represent the possible semantics of a composite event. Then, they applied a union aggregation technique to propagate properties of atomic events to composite events, and defined the attributes of composite events as a function of their sub-events. For example, Figure 23 shows two events e_1 and e_2 are composed via spatial and temporal operators to form composite event E .

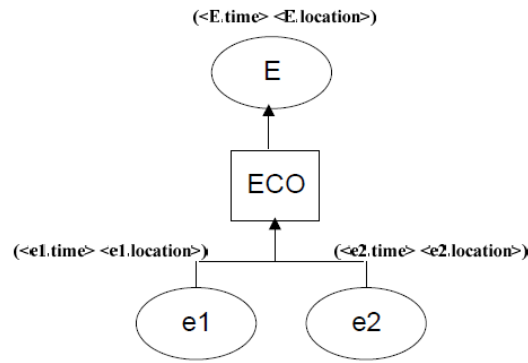


Figure 23: Illustration of atomic and composite events in [108].

In order to adequately provide the framework for composing events, the authors construct two graphs: i) a *SpatialGraph* using the RCC-8 spatial relationships [114], and ii) a *TemporalGraph* using Allen’s temporal relationship. Nevertheless, this work does not consider the social aspect of events (e.g., social relationships).

In a recent extension of their work [115], the same authors defined a generic ontological model, built on an event ontology, to infer sub events and their characteristics from personal photos using image metadata. This model is used to describe the vocabulary of a general domain event (such as trip) based on the *E* model* [116], which is an RDF-based data model to represent spatial, temporal and thematic relationships between data objects. Many properties are found in this semantic event model representing general relationships such as composition with *subeventOf*, similarity with *same-as*, spatiotemporal relationships like *occurs-during*, *occurs-at*, *co-occurring-with*, and *co-located-with*, etc. The authors also constructed a semantic language to model different types of entity properties and relationships related to an event domain on the basis of OWL. They added context information from heterogeneous data sources in order to provide very flexible models for high-level semantics of events, such as complex temporal formulas and spatial constraints and functions that cannot be expressed in OWL.

A number of studies have demonstrated the usefulness of social media in collecting data for emergency management [117-122]. For instance, the authors of [107] focus on real large-scale events (e.g., floods), which are important in emergency management. This study is useful for providing general overview information about events during a disaster or for after-the-fact analysis for training purposes. The authors presented a framework that allows the analysis of multimedia data using metadata (e.g., tags and title) associated with content found on social media platforms (Figure 24). They focus on identifying sub-clusters, which point to different hot spots in the emergency region. Subevents are defined here as events during a disaster which are separated from other events w.r.t. time or location. The first step of the framework contains a pre-selection of data (e.g., images, videos) from different repositories using user-supplied keywords, and filtered based on a date period inserted by the user. The second step is the clustering of Flickr photos and Youtube⁷⁵ videos into clusters representing sub-events, based on a *Self-Organizing Map*⁷⁶ (SOM). SOM takes as input a subset of relevant words computed based on the *tf-idf* scores. Results of this work [107] showed the suitability of their clustering algorithm for detecting sub-events. However, this is an offline approach where the optimal terms are selected before clustering. Therefore, its drawback is that it cannot be directly deployed in real-time during an emergency, and does not take geo-referenced data or date/time information into account.

⁷⁵ <https://www.youtube.com/?gl=FR>

⁷⁶ http://en.wikipedia.org/wiki/Self-organizing_map

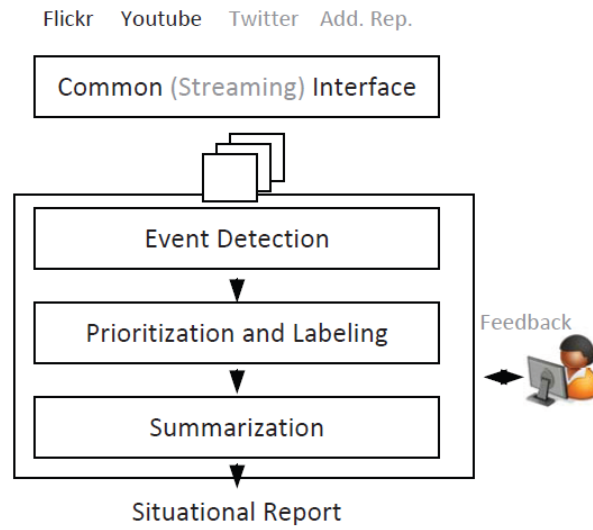



















Figure 24: Multimedia (metadata) exploration framework presented in [107].

In a later work, the same authors suggested a clustering approach to detect crisis-related sub-events online in real-time [123]. They have addressed the problem by referring to the topic detection and tracking (TDT) [67]. Events are similar to topics (and therefore sub-events to sub-topics), but different in that crisis-related events can evolve quite easily. Thus, the authors extended their first approach and explicitly included data containing geo-referenced information (i.e., longitude and latitude coordinates) to assist disaster management with Twitter, Flickr and YouTube data in real-time. Their approach relies on an online/incremental feature selection mechanism combined with an online clustering algorithm. The feature selection is dynamic, as a weighting function is used to calculate the importance of a term over different periods. It is based on the *learning and forgetting* model. The clustering algorithm is adapted to the Growing Gaussian Mixture Models (2G2M) algorithm [124]. The problem of this approach with respect to our goal of identifying relationships between events is that it identifies one specific type of events: *sub-events*. It has a spatial-temporal focus on the data but does not identify the spatial-temporal relations between events and sub-events.

IV.3.1 Discussion

In Table 5, we summarize the current event linking approaches and compare their characteristics.

Table 5: Comparison of event relations approaches.

Approaches	Spatial relations	Temporal relations	Social relations	Complex relations	Ontological events	User-defined relations
Eventory [109]						
DROP [81]						
EDF [106]						
EDF ^E [113]						
ECO [108]						
Event ontology [115]						
Crisis related sub-events [107]						

Social networks have become a large repository for detecting events from real world through user's published data. Previously, different solutions have been suggested to detect events [73, 78, 82, 83, 91, 105, 125] and few of them try to discover relationships among events [81, 106-109, 115, 123]. We believe that metadata contained in multimedia data, such as photos, could be of great importance for the event-linking task.

The main problems of these solutions with respect to our goal of computing links between events can be summarized as follows:

- **Focus on spatio-temporal relations:** Existing approaches do not account for all aspects of an event. Indeed, most of them focus on temporal relationships. Spatial relationships are sometimes considered. Nevertheless none of these approaches expresses the relations between events on the basis of the type of the relationships that link the event's participants.
- **Lack of granularity:** Most of these approaches do not incorporate granularity values for the temporal and spatial relations that could be identified between events.
- **Capability of detecting complex events:** Some of existing approaches allow the identification of complex events, especially sub-events. They are based on ontologies. Therefore, they do not allow the relations among events to be known if the ontology has not first been defined.
- **Possibility of discovering semantic relations:** Current approaches do not consider semantic relations at all. For one approach only [109], the system includes semantic constraints. This was done to allow the user to define his own semantic relationships.

V. Conclusion

Many social networks users are privacy conscious and may choose to not disclose all information about them online. However, an adversary is perfectly able to use public information shared by their contacts or embedded in multimedia content to discover hidden information on their profiles. It is therefore crucial to provide solutions for enriching the user profile.

So far, many approaches examine the potential of identifying a person using the Web. But they suppose that the person has a priori related information on the Internet. Inputs, queries and results are text-based. Unlike textual attributes, multimedia content cannot be approached without special processing to reduce uncertain decisions that overcome when similarity operators come to play. Photos shared within online social networks are useful for obtaining

better background knowledge needed for preserving individuals' privacy. Moreover, we believe that investigating events and discovering link types between events will serve as a base for future applications that require semantically enriched profiles. In this study, we address these challenges and present our approach for detecting and linking events using photos shared within online social networks.

Chapter 3: Privacy-Preserving Framework for Multimedia Sharing

Abstract

Multimedia documents' publication in general and photo sharing in particular have become part of the routine activity of many individuals and companies. Such data sharing puts at risk the privacy of individuals, whose identities need to be kept secret, when adversaries get the ability to associate the multimedia document's content to possible trail of information left behind by the individual. In this chapter, we propose *de-linkability*, a privacy-preserving constraint to bound the amount of information shared that can be used to re-identify individuals. We provide a sanitizing *MD* *-algorithm to enforce *de-linkability* along with a utility function to evaluate the utility of multimedia documents that is preserved after the sanitizing process. A set of experiments are elaborated to demonstrate the efficiency of our technique.

I. Introduction

The publication of personal information on social media and blogging sites has resulted in the users' exposure to privacy breaches. In essence, preserving anonymity is becoming increasingly challenging with the tools and information available on the Internet today. It goes without saying that every anonymous multimedia document published can be put at risk and linked back to the individual without appropriate anonymization techniques.

As mentioned in the previous chapter, most of the works done in the literature to preserve anonymity focus on structured relational data [29, 126] while the only few techniques [38, 39] proposed to handle identity anonymization in multimedia documents assume textual data with no reference whatsoever to multimedia objects such as images and videos.

In this chapter, we provide a generic framework and efficient algorithm for dealing with any kind of multimedia documents. We propose *de-linkability*, a novel technique for preserving individual privacy before sharing multimedia documents. *De-linkability* ensures that individuals' identifiable information composed of both textual and multimedia content cannot be used to infer his/her identity. An evaluation based on a real image dataset demonstrates that our privacy-preserving constraint is able to bound the amount of information that can be used to re-identify the individual.

The main contributions of this chapter can be summarized as follows:

- We formally define the identity anonymization problem in multimedia documents composed of textual and multimedia content.
- We quantify the re-identification threat which is highly dependent on how much information can be acquired from i) adversaries' background knowledge and ii) external sources containing relevant information related to the anonymized individual.

- We present our sanitizing MD^* -algorithm that allows to sanitize multimedia documents' content and preserve at the same time their utility in order to achieve the *de*-linkability.
- We provide a utility measure to determine to which extent a multimedia document remains consistent after the sanitizing process.

The remainder of this chapter is organized as follows. In Section 2, we present the adversary model adopted in our study. Our data model definitions and operators are presented in Section 3. In Section 4, we give a formal definition of the re-identification problem. Section 5 is dedicated to present the *de*-linkability privacy constraint and to show how it is possible to preserve individual anonymity using a multimedia document sanitizing algorithm (the MD^* -algorithm) and a utility measure. In Section 6, we evaluate our sanitizing algorithm to finally conclude and discuss our future research directions in Section 7.

II. Adversary Model

In our adversary model, we assume that the adversary, that we call *cyberstalker*, knows that a given individual, that we call *cyberstalkee*⁷⁷, is hiding his/her identity (e.g., François Fillon in the Motivation Scenario in Chapter 1). We also assume that the *cyberstalker* has access to public information enabling him/her to link some personally identifying information, in a shared multimedia document, to the *cyberstalkee*. More subtle, we assume that the *cyberstalker* has no prior knowledge of specific values for the stalked individuals. For example, the *cyberstalker* described in our motivating example does not know a-priori that “Château de Beaucé” is the residence of the *cyberstalkee*.

⁷⁷ Both terms *cyberstalkee* and *individual* will be used interchangeably in the remainder of this chapter.

III. Data Model

We start by defining the data model and the basic notations (Table 6) used in the remainder of this chapter.

Table 6: Notations used in our approach.

u	an individual with anonymized identity
pf_u	an individual profile
mo	a multimedia object
MD_u	a multimedia document of u to be sanitized
MD_β	a multimedia document publicly accessible to adversaries extracted from an external source E
S_{MD}	a multimedia document signature
E	an external source such as the social website, domain specific database, etc.
A	a set of attributes relevant to the multimedia document content
α	an association threshold
β	an identification threshold
ϖ	an aggregation function such as average, minimum, max, etc.
R_w	words relevance
R_m	multimedia relevance
U	multimedia document utility

III.1 Data Definition

Definition 1 – Attribute Set: A is a set of attributes where $\forall a_i \in A$ for $(1 \leq i \leq |A|)$, a_i can be any attribute of the Dublin core metadata element set⁷⁸ such that $\{source, description, date,$

⁷⁸ <http://dublincore.org/>

contributor, format or the MPEG-7 semantic set⁷⁹ $\{semantic_place, concept, state, event, object\}$ or any domain specific attribute (e.g., spatial or temporal domain). We use $ma_i \in A$ to denote a multimedia attribute whose values are of complex structure such as a BFILE/BLOB, an URL/URI, an URL/URI augmented with a primitive to represent a salient object (e.g., Minimum Bounding Rectangle, Circle) or a multimedia object.

Definition 2 – Multimedia Object: Let mo be any type of multimedia data such as an image, a video, or a salient object describing an object of interest (e.g., face of a person.). mo is formally represented as:

$$mo: < A_m, V, O, MO, \varsigma >$$

where:

- $A_m \subseteq A$ is a subset of attributes of A whose values are used to identify a multimedia object mo .
- V is a set of values describing the multimedia object. $\forall v_i \in V$ for $(1 \leq i \leq |V|)$ $v_i \in D(a_j)$ where a_j is an attribute of A_m .
- O is the raw data of the multimedia object. $O \in D(a_i)$ where a_i is a multimedia attribute of A_m . $O(mo)$ denotes the raw data of multimedia object mo .
- MO is a set of multimedia objects contained in mo . In this work, we only consider image data. Therefore, we will not have MO as in the case of video which is constituted of scenes and frames.
- $\varsigma \subseteq A_m \times X = \{(a_j, x_i) | a_j \in A_m, x_i \in V \text{ or } x \text{ is } O\}$ is an association function that assigns each attribute a_j to its corresponding value which is either a textual $v_i \in V$ or a multimedia raw data O .

⁷⁹ <http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm>

For example, Figure 25 shows multimedia objects mo_{beauce} and mo_{manoir} representing two images of “Château de Beaucé” where “keywords” is an attribute of mo , O contains the raw data and MO is the empty set of multimedia objects contained in mo .

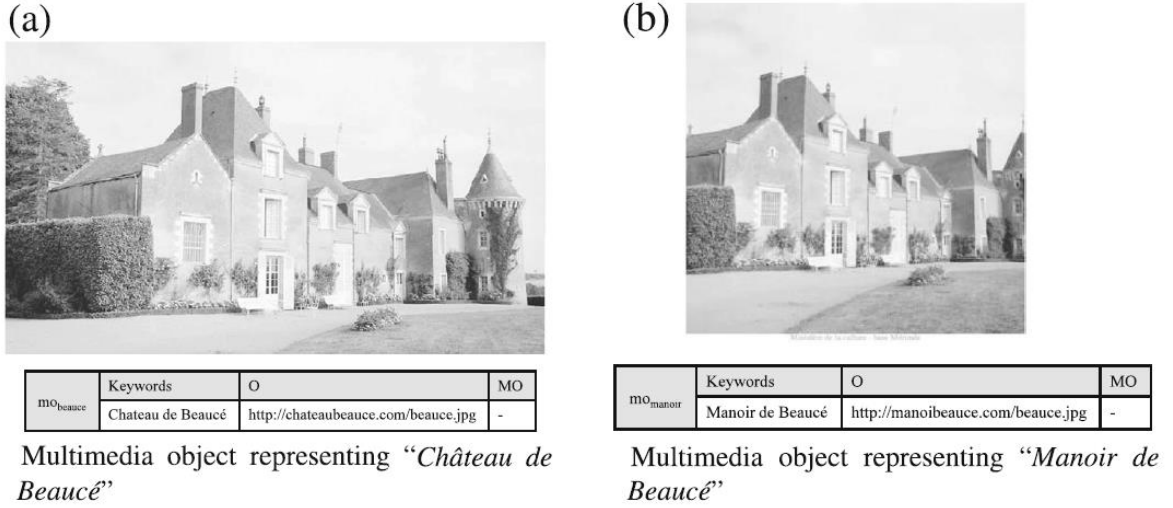


Figure 25: Examples of typical image descriptions using our multimedia object representation. Figure 1(a) and Figure 1(b) show similar content, contained in two different multimedia documents.

Definition 3 – Individual Profile: Let u be a *cyberstalkee*, we denote by pf_u the profile of u formally defined as:

$$pf_u: \langle A_p, PI, MO, \gamma \rangle$$

where:

- $A_p \subseteq A$ is a subset of attributes of A whose values are used to identify an individual profile pf_u .
- PI is a set of values describing the individual’s personal information. $\forall v_i \in PI$ for $(1 \leq i \leq |PI|)$, $v_i \in D(a_j)$ where a_j is an attribute of A_p .

- MO is a set of multimedia objects attributed to u such that $\forall mo_i \in MO$ for $(1 \leq i \leq |MO|)$ $mo_i \in D(ma_j)$ where ma_j is a multimedia attribute $\in A_p$. In this work, we only consider the set of images related to u .
- $\gamma \subseteq A_p \times X = \{(a_j, x_i) | a_j \in A_p, x_i \in PI \text{ or } x_i \in MO\}$ is an association function that assigns each attribute a_j to its corresponding value which is either a textual $v_i \in PI$ or multimedia $mo_i \in MO$.

Referring back to our scenario, a typical profile of the previous french Prime Minister François Fillon would be:

$pf_{Fillon} : ((name, \textit{François Fillon}), (job, \textit{Prime Minister}), (country, \textit{France}), (email, \textit{fcafillon@wanadoo.fr}), (home, mo_{beauce}))$

Definition 4 – Multimedia Document: Let MD be a multimedia document. MD is two dimensional and composed of a set of words and multimedia objects. It is formally defined as follows:

$$MD : \langle W, mo \rangle$$

where:

- W is a base text that represents the document's content where $\forall w_i \in W$ for $(1 \leq i \leq |W|)$, w_i is a word contained in MD or a metadata (attribute information) of the document.
- MO is a set of multimedia objects where $\forall mo_i \in MO$ for $(1 \leq i \leq |MO|)$, mo_i is a multimedia object contained in MD .

An example of a multimedia document could be, but not limited to, personal blogs, set of tweets, newspaper articles, etc. Typically, these documents are composed of words and multimedia objects.

Now that we have defined our multimedia document, we present in the following what we call a multimedia document signature (S_{MD}).

Definition 5 – Multimedia Document Signature: Let MD be a multimedia document, a multimedia document signature denoted by S_{MD} is a subset of MD composed of textual and multimedia content. S_{MD} is created using $S_{MD} = IA(MD, A_s)$ where IA is a function used to retrieve from MD relevant words and multimedia objects related to the subset of attributes $A_s \subseteq A$.

We assume that not all attributes found in a multimedia object provide meaningful clues that could lead to re-identify the *cyberstalkee*. In essence, the idea is to generate signatures that are mainly related to the individuals. This could be done by determining the most significant and relevant attributes retrieved from the document's content and/or using some of the attributes from the individuals' profiles. These attributes help reduce the error rate of individual name disambiguation [127], particularly when the individual's profile is considered as a relevant source of attributes. For instance, it is unlikely for an individual working in a Health Care Department to be related to Computer Science. In other terms, some of the words and multimedia objects should more likely be related to the medical field instead of computing.

The followings are three sample multimedia documents' signatures generated based on the attributes *Country*, *Event*, *Location* and *Image*. $S_{MD_{Fillon}}$ is the anonymous multimedia document signature of Prime Minister François Fillon.

$$S_{MD_{Fillon}} : ((\text{visiting}, \text{"Japan"}), (\text{visit}, \text{"Meeting"}), (\text{annotation}, \text{"@beauce"}), (\text{home}, \text{mobeauce}))$$

Both $S_{MD_{\beta_1}}$ and $S_{MD_{\beta_2}}$ are publicly available multimedia documents signatures related to Prime Minister *François Fillon*.

$$S_{MD_{\beta_1}} : ((\text{name}, \text{"François Fillon"}), (\text{country}, \text{"France"}), (\text{visiting}, \text{"Japan"}), (\text{visit}, \text{Meeting}))$$

$$S_{MD_{\beta_2}} : ((\text{name}, \text{"Francis Fillon"}), (\text{annotation}, \text{"Home"}), (\text{home}, \text{mOmanoir}))$$

III.2 Data Comparison

We provide, in this section, the appropriate operators to address both multimedia and textual content of multimedia documents.

Definition 6 – Estimated Equality: Let W_1, W_2 be two sets of words over which an association function f can be used. Their estimated equality is computed as follows:

$$equ(W_1, W_2) = \varpi(f(w_w^1, w_1^2), \dots, f(w_m^1, w_r^2)) \rightarrow [0, 1]$$

where:

- w_i^1, w_j^2 are two words of W_1 and W_2 respectively, $m = |W_1|$ and $r = |W_2|$.
- f is an association function defined as:

$$f(w_i^1, w_j^2) = \begin{cases} 1 & \text{if } w_i^1 \in W_1 \text{ is the same as } w_j^2 \in W_2 \\ 0 & \text{otherwise} \end{cases}$$

- ϖ is an aggregation function (e.g., max, min, avg, etc.) used to aggregate association functions' scores.

In our example, the estimated equality takes, for instance, the set of words in the Wikipedia article about François Fillon and compares them with the set of words on his Twitter page. For instance, if we use the substring function SUBSTR⁸⁰ between words from the two sets, we see that *beauce* (from the Wikipedia page⁸¹) is the substring value we wish to find from *fdebeauce* (from the Twitter page).

The estimated equality is used to identify the amount of common textual values found in multimedia documents (or any subset of them). Alternatively, multimedia documents contain complex types such as images and videos which cannot be approached using traditional

⁸⁰ SUBSTRING (expression, start, length) = SUBSTR(*fdebeauce*, 4, 6) = *beauce*

⁸¹ http://en.wikipedia.org/wiki/Francois_Fillon

equality operators. We define in the following, an estimated similarity operator to process multimedia objects.

Definition 7 – Estimated Similarity: Let MO_1, MO_2 be two sets of multimedia objects over which n similarity functions s_1, \dots, s_n can be used. Their similarity score is computed as follows:

$$sim(MO_1, MO_2) = \varpi(s_1(mo_1^1, mo_1^2), \dots, s_n(mo_m^1, mo_r^2)) \rightarrow [0, 1]$$

where:

- mo_i^1, mo_j^2 are two multimedia objects of MO_1 and MO_2 respectively where $m = |MO_1|$ and $r = |MO_2|$.
- s_k is a *unit* similarity function comparing multimedia objects $mo_i^1 \in MO_1$ and $mo_j^2 \in MO_2$. We note that $s_k(mo_i^1, mo_j^2)$ compares mo_i^1 and mo_j^2 based on their attributes and raw data. s_k returns a score between $[0, 1]$, where 0 expresses a total divergence and 1 a complete similarity.
- ϖ is an aggregation function used to aggregate the computed similarity scores.

We give below an example to illustrate the estimated similarity between two sets of multimedia objects. Therefore, we propose a unit similarity function⁸² that takes coordinates of two images and returns 1 if the input coordinates belong to the same geographical location. We see in our example that one of the images published on the Twitter account of François Fillon was captured in the same geographical coordinates (*Latitude* : 48.357483, *Longitude* : -1.116662) as the image of “*Château de Beaucé*” found on the Wikipedia page of François Fillon.

⁸² For example, $compareGPS((lat_1, long_1), (lat_2, long_2)) = 1$ if $(ConvertFromLatLong(lat_1, long_1) = ConvertFromLatLong(lat_2, long_2))$; 0 otherwise

Definition 8 – Cross-Matching Score: Let S_{MD_1} , S_{MD_2} be the signatures of two distinct multimedia documents. The cross-matching score between their components (W and MO) is computed as follows:

$$match(S_{MD_1}, S_{MD_2}) = \lambda_m \times f(W_1, MO_2) + (1 - \lambda_m) \times f(W_2, MO_1) \rightarrow [0, 1]$$

where:

- $f(W_1, MO_2)$ and $f(W_2, MO_1)$ are association functions to determine the association of a set of words contained in S_{MD_1} (S_{MD_2} respectively) with the set of multimedia objects contained in S_{MD_2} (S_{MD_1} respectively). $f(W_1, MO_2)$ is defined as:

$$f(W_1, MO_2) = \varpi(f_1(w_1^1, mo_1^2), \dots, f_n(w_n^1, mo_r^2)) \rightarrow [0, 1]$$

where:

- f_k is a *unit* association function used to determine whether there is an association between a word $w_1 \in W_1$ with a multimedia object $mo_1 \in MO_2$.

We note that $f_1(w_i^1, mo_j^2)$ determines the association between w_i^1 and mo_j^2 based on the attributes and values of the latter. For example, it may associate a GPS coordinates, specified as one of the multimedia objects' metadata, with a word representing the corresponding location. The function f_k returns a score between $[0, 1]$, where 1 represents a perfect association of the attributes and 0 represents the absence of association.

- ϖ is an aggregation function used to aggregate the computed association scores.
- $\lambda_m \in [0, 1]$ allows to assign priorities to $f(W_1, MO_2)$ and $f(W_2, MO_1)$, based on the relevance of multimedia objects in the multimedia documents.

To illustrate the influence of the cross-matching, we take as input the set of words from the Wikipedia page about François Fillon and the set of multimedia objects from his Twitter page. We define a function⁸³ that maps words from the first set to geographical coordinates in the second set. By doing so in our example, we found that the profile image published on the Twitter account and “*Château de Beaucé*” found on the Wikipedia page refer to the same location.

We show in the following how multimedia documents intersection can be determined using selective intersection.

Definition 9 – Selective Intersection: Let S_{MD_1} , S_{MD_2} be the signatures of two distinct multimedia documents. Their selective intersection is defined as:

$$S_{elInt}(S_{MD_1}, S_{MD_2}) = \left\| \sum_{a_i} wa_i \times equ_{a_i}(W_1, W_2) + \sum_{a_j} wa_j \times sim_{a_j}(MO_1, MO_2) + \sum_{a_k} wa_k \times match_{a_k}(S_{MD_1}, S_{MD_2}) \right\|$$

where:

- a represents an attribute for which an equality, similarity and/or cross-matching score should be computed. Such attributes, defined in the attribute set, can be used to selectively choose relevant content in multimedia documents. For instance, it is

⁸³ For example, $compareLocations((lat_1, long_1), (address_2)) = 1$ if $(ConvertFromLatLong(lat_1, long_1) = address_2)$; 0 otherwise

possible to capture the amount of common information related to the attribute *Person*. This refers to computing equality, similarity and cross-matching of words and multimedia objects that are related to this attribute for both multimedia documents' signatures S_{MD_1}, S_{MD_2} .

- wa is the weight assigned to attribute a where its magnitude depends on the normalizing assumptions.

Selective intersection returns a normalized score $\in [0, 1]$ computed based on equality, similarity and cross-matching of multimedia documents content. For instance, let us compute the selective intersection between $S_{MD_{Fillon}}$ and both $S_{MD_{\beta_1}}$ and $S_{MD_{\beta_2}}$. We adopt the *max* aggregation function to compute the equality, similarity and/or matching scores for each attribute and finally determine their average score. The selective intersection based on the attributes *Country*, *Event* and *Location* is detailed below:

$$S_{el}I_{nt}(S_{MD_{Fillon}}, S_{MD_{\beta_1}}) = \frac{\frac{1+1+0}{3} + 0 + 0.3}{3} = 0.32$$

$$S_{el}I_{nt}(S_{MD_{Fillon}}, S_{MD_{\beta_2}}) = \frac{\frac{0}{3} + 0.8 + 0.5}{3} = 0.44$$

We assume, in this case, that the estimated similarity between multimedia objects mo_{beauce} and mo_{manoir} in $S_{MD_{Fillon}}$ and $S_{MD_{\beta_2}}$ returns a 0.8 score. The cross-matching score between multimedia object mo_{beauce} and *France* in $S_{MD_{Fillon}}$ and $S_{MD_{\beta_1}}$ respectively returns a 0.3 score for semantic similarity. This matching returns a 0.5 score between $S_{MD_{Fillon}}$ and $S_{MD_{\beta_2}}$ based on the matching between “keywords” attribute of the multimedia object (mo_{manoir}) and *@beauce*.

Unlike mutual information metric [128] which is based on joint probability measures, our selective intersection is a non-correlation based metric where the count of each value in the signatures has minimum influence on the overall computation score. Specifically and for

privacy reasons, this assumption is useful to determine the “minimum” intersection between multimedia documents where weighted attributes reflect relevant association measure if processed efficiently. We will show in the following definition, the premise of multimedia documents’ association.

Definition 10 – α -association: Let MD_1, MD_2 be two distinct multimedia documents. We say that an α -association exists between MD_1 and MD_2 if their selective intersection $S_{elInt}(S_{MD_1}, S_{MD_2})$ is greater than α where:

- S_{MD_1} and S_{MD_2} represent corresponding multimedia documents signatures.
- $\alpha \in [0, 1]$ is the association threshold.

α -association expresses the presence of a possible association between two multimedia documents represented by their signatures. It measures the strength of an association between two multimedia document signatures based on their common information composed of both textual and multimedia content.

Suppose we set the threshold $\alpha = 0.4$. Since $S_{elInt}(S_{MD_{Fillon}}, S_{MD_{\beta_1}}) < 0.4$, MD_{Fillon} when combined with MD_{β_1} does not break the anonymity of François Fillon. However, $S_{elInt}(S_{MD_{Fillon}}, S_{MD_{\beta_2}}) > 0.4$ can put at risk the privacy of François Fillon and hence allows his identification.

III.3 Identity Anonymization Problem

In the presence of adversaries with sophisticated tracking abilities, privacy and ownership preserving of shared data tends to be a complex task. Such adversaries, armed with plausible background knowledge and a wide range of accessible web-based social information, compromise anonymization techniques and put at risk individuals’ privacy. Here, we express

the identity anonymization problem that could arise when sharing or publishing multimedia documents as the amount of information accessible by the adversary and that can be, at the same time, associated with the owner of the published multimedia documents. It is formally defined as follows:

Definition 11 – Identity Anonymization Problem: Let MD_u be the multimedia document of an individual u . There are two thresholds α and β . We say that an adversary is able to re-identify u from MD_u if $\exists MD_\beta$, a *publicly available* multimedia document, such that:

- MD_u and MD_β are α -associated and,
- The knowledge related to u that can be obtained from MD_β is greater than β . It is expressed as a β -association between MD_β and the individual profile pf_u where α is the association threshold, β is an identification threshold and both α and β are predefined according to the specific application environment.

It is difficult to know how much the adversaries know and to what extent their ability to disclose individuals' identities can be compromising. Here, we only avoid leaking information to the *cyberstalker* except for what he/she already has. Such assumption is not different than the one adopted by differential privacy [126] where our main objective is essentially providing constraints on the release of the data. Differential privacy provides a model for privacy-preserving analysis of statistical databases, which are collections of records which contain statistical information about individuals. It is characterized by a property of algorithms operating on the data, typically computing some statistical function of the data.

IV. Privacy Preservation Prior to Publication

Preserving privacy requires that the *cyberstalker* remains unable to detect the anonymized identity of the *cyberstalkee*, owner of the multimedia document to be published. As we have stated in the previous section, a re-identification threat occurs mainly due to:

- the link between his/her related multimedia document MD_u and a multimedia document MD_β accessible by the *cyberstalker*, and
- the amount of information extracted from MD_β and associated with u .

Controlling the latter can be, on one hand, a burden or eventually unrealizable due to accessibility issues while, on the other hand, breaking the link between multimedia documents is achievable and can be done using *de-linkability*.

de-linkability. Given a cyberstalkee u and a multimedia document MD_u , the *de-linkability* privacy-preserving constraint is satisfied if $\forall MD_\beta \in \sigma_{E_u}(E)$ that is β -associated with pf_u , MD_u cannot be linked to MD_β through an α -association, where $\sigma_{E_u}(E)$ is a selection on an external source ε based on a conjunctive set of words and/or multimedia objects (E_u) related to u .

de-linkability breaks the link between a multimedia document (to be published) and any other document accessible to a cyberstalker and that can be linked to u . It is important to note that the content of E_u that is used to retrieve multimedia documents MD_β from the external source should be considered carefully in order to reduce the scope of potential error. A straightforward assumption is to consider this content as a subset of the individual's profile including both identifying and quasi-identifying values.

IV.1 Achieving *de-linkability*

de-linkability can be achieved in textual documents in a straightforward way using extension of traditional anonymization techniques such as suppression, substitution or generalization relationships between domains and values [29, 36, 37] for textual values in MD_u as long as there is no MD_β that can be α -associated with MD_u . Unsurprisingly, multimedia objects need a special interest. Eventually, the objective is to break linkable objects that could contribute in re-identifying the anonymized individual. More subtle is to hide and/or disseminate

multimedia objects content while at the same time preserving a minimum semantic or visual coherence. In this study, we do not provide an in-depth details on how multimedia objects content could be protected. This matter is left for future work. We only use traditional techniques to protect salient objects as in [129] where the authors protect textual and image data through flexible low-level adapted security rules, while in [130] object substitution is adopted. In [131], blurring proved efficiency, and objects removal from images and videos were addressed in [132-137]. Here, we refer to this process as document sanitization which we formally define as follows:

Definition 12 – Multimedia Document Sanitization: Let MD_u be the multimedia document related to a *cyberstalkee* u . Given \tilde{G}_W and \tilde{G}_{MO} , two corresponding sanitizing functions, we say that MD_u is sanitized, denoted by $MD_u^* = \tilde{G}_{(W,MO)}(MD_u)$ if both words and multimedia objects are sanitized $\tilde{G}_W(W_{MD_u})$ and $\tilde{G}_{MO}(MO_{MD_u})$.

Multimedia document sanitization ensures that the specified content (W,MO) is either removed, suppressed, generalized and/or protected in the multimedia document MD_u based on the sanitization function.

IV.2 Multimedia Document Sanitization: MD^* – algorithm

MD^* -algorithm is used to sanitize a multimedia document and protect the *cyberstalkee*'s identity. As mentioned in the pseudo-code, the algorithm takes a multimedia document MD_u , a set of attributes A_s (used to extract multimedia document signature), the *cyberstalkee* profile pf_u along with E_u and both association and identification thresholds α, β . It returns a sanitized multimedia document (MD_u^*).

Algorithm 1 \mathcal{MD}^* -algorithm

Require: a multimedia document \mathcal{MD}_u , set of attributes \mathcal{A}_s over \mathcal{MD}_u , an individual profile pf_u , conjunctive set of words and/or multimedia objects E_u , association threshold α and identification threshold β

Ensure: Multimedia Document Sanitization \mathcal{MD}_u^*

```

1:  $S_{\mathcal{MD}_u} = IA(\mathcal{MD}_u, \mathcal{A}_s)$                                 ▷ Generate Multimedia Signature on  $\mathcal{MD}_u$ 
2: for each  $\mathcal{MD}_\beta$  in  $\sigma_{E_u}(E)$  do
3:    $S_{\mathcal{MD}_\beta} = IA(\mathcal{MD}_\beta, \mathcal{A}_s)$                                 ▷ Generate Multimedia Signature on  $\mathcal{MD}_\beta$ 
4:   if  $S_{elInt}(S_{\mathcal{MD}_\beta}, pf_u) > \beta$  then
5:     while  $S_{elInt}(S_{\mathcal{MD}_u}, S_{\mathcal{MD}_\beta}) > \alpha$  do
6:       Retrieve least significantly threatening  $W_\beta$  and  $MO_\beta$ 
7:        $\mathcal{MD}_u^* \leftarrow \tilde{G}_{(W_\beta, MO_\beta)}(\mathcal{MD}_u)$                 ▷ Sanitize  $\mathcal{MD}_u$  based on  $W_\beta$  and  $MO_\beta$ 
8:     end while
9:   end if
10: end for
    
```

The MD^* -algorithm extracts in Step 1 the multimedia document signature S_{MD_u} using the extraction function IA . It sanitizes MD_u from Step 2 to 10.

In Step 3, it extracts the signature of a multimedia document MD_β retrieved from an external source E based on the set of entities E_u related to u . In order to determine the amount of information related to u and that can be obtained from MD_β , we compute the selective intersection on MD_β and the *cyberstalkee* profile pf_u . If their selective intersection $S_{elInt}(S_{MD_\beta}, pf_u)$ is greater than β , the link between MD_u and MD_β should be anonymized as done from Step 5 to 8. That is, as long as they are α -associated, the least⁸⁴ significant W_β and MO_β are sanitized in MD_u .

⁸⁴ The importance of retrieved W_β and MO_β is determined based on the priority thresholds prefixed in the selective intersection function.

IV.3 Utility Estimation

To ensure safety, there is trade-off to be made at the stake of utility in order to meet strong privacy requirements. While this could be limited in general, it is considered an absolute necessity in order to establish trust between data owners and data providers. This issue has been the essence of several works [138-140] that provide data anonymization. Here, we determine to what extent a multimedia document remains consistent after the sanitizing process. In particular, we provide an estimation of utility based on the relevance of both words and multimedia objects sanitized.

Definition 13 – Words Relevance: Let W^* be the set of words sanitized from MD_u , we define words relevance, denoted by $R_w(W^*)$, as the raw frequency of words addressed by the sanitizing process. It is computed as follows:

$$R_w(W^*) = \sum_{w_i \in W^*} \frac{c(w_i)}{N_w}$$

where:

- W^* is the set of sanitized words in MD_u^* ,
- $c(w_i)$ is the number of times w_i appearing in the multimedia document,
- N_w is the total number of words in the multimedia document.

Note that R_w assigns weights to individual words sanitized from the multimedia document. It determines the relevance despite the adopted anonymization technique (generalization, suppression or encryption).

Unlike words, determining the relevance of multimedia objects depends on the raw data of the multimedia object.

Definition 14 – Multimedia Relevance: Let MO^* be the set of multimedia objects sanitized in MD_u , we define multimedia relevance, denoted by $R_m(MO^*)$, as the importance of multimedia

objects sanitized from the multimedia document MD_u . $R_m(MO^*)$ is computed based on multimedia objects raw data. It is determined as follows:

$$R_m(MO^*) = \frac{\sum_{(mo_i \in MO^*)} r_m(O(mo_i))}{\sum_{(mo_i \in MO^*)} \rho_i}$$

where:

- MO^* is the set of sanitized multimedia objects in MD_u^* ,
- $r_m(O(mo_i)) = \rho_i \frac{sizeOf(O(mo_i))}{sizeOf(O(mo_j))}$ is the relevance of the raw data of mo_i ,
- ρ_i is the importance threshold of the multimedia object mo_i . It can be computed based on the association of the raw data of mo_i with words and/or multimedia objects from the individual's profile,
- $sizeOf(O(mo_i))$ is the size of the raw data of multimedia object mo_i in terms of width and height,
- $sizeOf(O(mo_j))$ is the size of the *container* mo_j where $mo_i \in MO(mo_j)$.

We provide in the following a formal definition of the utility of a multimedia document.

Definition 15 – Multimedia Document Utility: Let MD_u^* be a sanitized multimedia document of an individual u , we denote by $U(MD_u^*)$ the utility measure of MD_u^* which is the estimated coherence given the relevance of sanitized words and multimedia objects from MD_u . It is formally defined as follows:

$$U(MD_u^*) = \frac{1 - R_w(W^*) \times R_m(MO^*)}{1 + R_w(W^*) \times R_m(MO^*)}$$

where $R_w(W^*)$ and $R_m(MO^*)$ are word and multimedia relevance metrics.

U is used to express the trade-off between privacy and utility. It shows at which point a sanitized multimedia document can be considered useless according to the amount of relevant information it contains.

V. Experiments

In this section, we present a set of experiments to evaluate the efficiency of our approach. We implemented the MD^* -algorithm code in Java and conducted experiments using a 3.4 GHz Intel Core i7 with 16 GB RAM.

V.1 Dataset configuration

We used 200 individuals of the dataset published⁸⁵ by the authors of [1]. For each individual, we grouped 100 of his/her tweets to form his/her MD_u . These MD_u have been filtered to remove identifying names. OpenCalais api⁸⁶ is used to extract attributes from multimedia documents MD_u and MD_β . We actually used the most relevant attributes extracted based on a predefined threshold that we have set to 0.5 (this threshold can be used to fine-tune the evaluation results and include relevant attributes).

We limited our use of multimedia objects to images. We specifically used the Zemanta api⁸⁷ to retrieve and associate images with their related words contained in MD_u . As a matter of fact, the images that were mainly retrieved from the web, compensate the lack of metadata that could be used to link words to their corresponding images. That being said, the use of the Zemanta api enriched the content of MD_u with multimedia objects that could be used to re-identify individuals.

⁸⁵ <http://wis.ewi.tudelft.nl/umap2011/>

⁸⁶ <http://www.opencalais.com/>

⁸⁷ <http://developer.zemanta.com/>

Individual profiles pf_u were downloaded using the Twitter api⁸⁸. For our assessment, we only focused on four profile attributes namely *name*, *screen name*, *location* and *profile image url*.

As per *cyberstalkee*, we retrieved up to 10 relevant multimedia documents MD_β using the Google api⁸⁹ applying to the individual *name* combined to relevant content from his/her related MD_u . This way, we can assert that the retrieved multimedia documents MD_β are related to the *cyberstalkee* at hand at least through their names.

To compare images, we used the phash function⁹⁰ and assigned a manual weight of 0.5 to the estimated similarity for the selective intersection $S_{el}I_{nt}$.

V.2 Evaluation Results

We elaborated a set of measurements to evaluate the efficiency of the MD^* -algorithm. These measurements can be summarized as follows:

- Evaluating the identity anonymization problem represented by the percentage of individuals re-identified;
- Determining the uncertainty raised after sanitizing multimedia documents;
- Evaluating the utility of multimedia documents after the sanitizing process;
- Determining the computational cost of our MD^* -algorithm.

V.2.1 Evaluating Privacy

In this first test, we evaluated the identity anonymization problem represented by the percentage of individuals identified according to what they have published in their MD_u and

⁸⁸ <https://dev.twitter.com/>

⁸⁹ <https://dev.twitter.com/>

⁹⁰ <http://phash.org/docs/howto.html>

their related multimedia documents MD_β . We fixed the identification threshold $\frac{1}{\beta} = 10$ in order to capture a significant number of multimedia documents related to individual u and used various association thresholds $\frac{1}{\alpha} = 2, 4, 6, 8$ and 10 . The results shown in Figure 26 show the percentage of re-identified individuals (in Figure 26 (a)) and the number of threatening MD_β (in Figure 26 (b)).

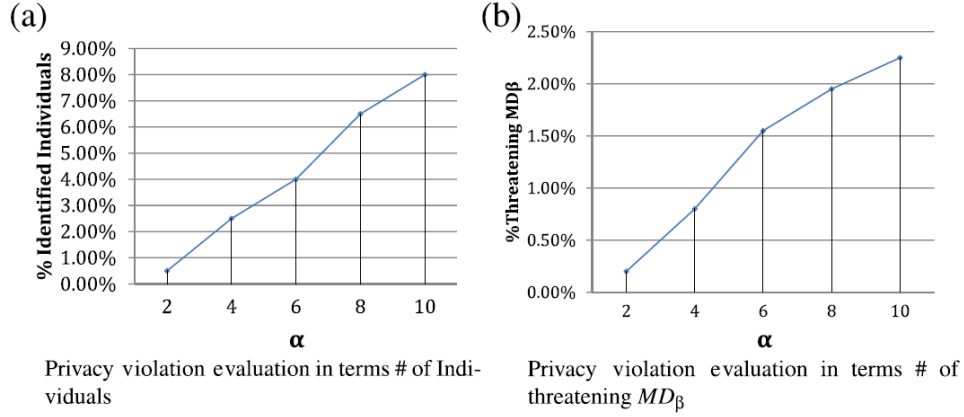


Figure 26: Privacy violation evaluation.

We can see that when the association threshold increases, there is a higher chance of linking individuals to the multimedia documents MD_β retrieved from the external source and eventually leading to their re-identification.

V.2.2 Evaluating Uncertainty

We evaluated the MD^* -algorithm to determine the increasing uncertainty raised due to the sanitizing process. To do so, we calculated the average entropy [141] of individuals' multimedia documents MD_u in a pre- and post-sanitizing process. As a matter of fact, for each individual's multimedia document, we computed its entropy based on the most relevant attributes used to generate its own multimedia document signature (see Definition 5) as:

$$Entropy(MD_u) = - \sum_{a \in A} \Pr(a) \log(\Pr(a))$$

where a is the related attribute. We estimate the uncertainty to be: $|Entropy(MD_u) - Entropy(MD_u^*)|$ where MD_u^* is the sanitized multimedia document. The results are shown in Figure 27.

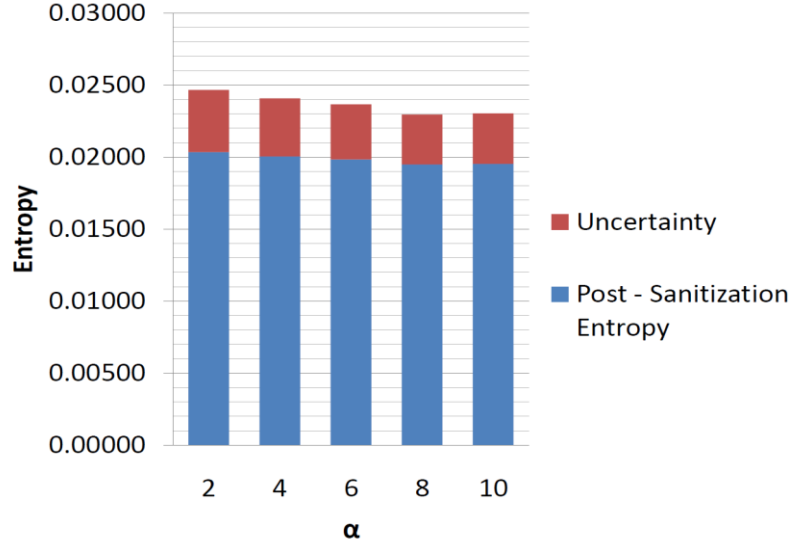


Figure 27: Uncertainty evaluation.

Figure 27 shows that the uncertainty caused by the sanitizing process is relatively small. This uncertainty could get even smaller if sanitizing multimedia objects was approached differently using blurring or pixelizing techniques which might preserve the semantic and coherence of images' content. This requires a more extensive study that we would like to address in the future.

V.2.3 Evaluating Utility

To evaluate the utility of multimedia documents that have been subject of a sanitization process, we sanitized a specific percentage (20, 40, 60 and 80%) of words and multimedia objects chosen randomly from the multimedia document signatures of 100 MD_u . Here, the salient objects which are represented using our multimedia object representation have been sanitized. The resulting utility computed in terms of words and multimedia relevance metrics of each of the multimedia documents is shown in Figure 28.

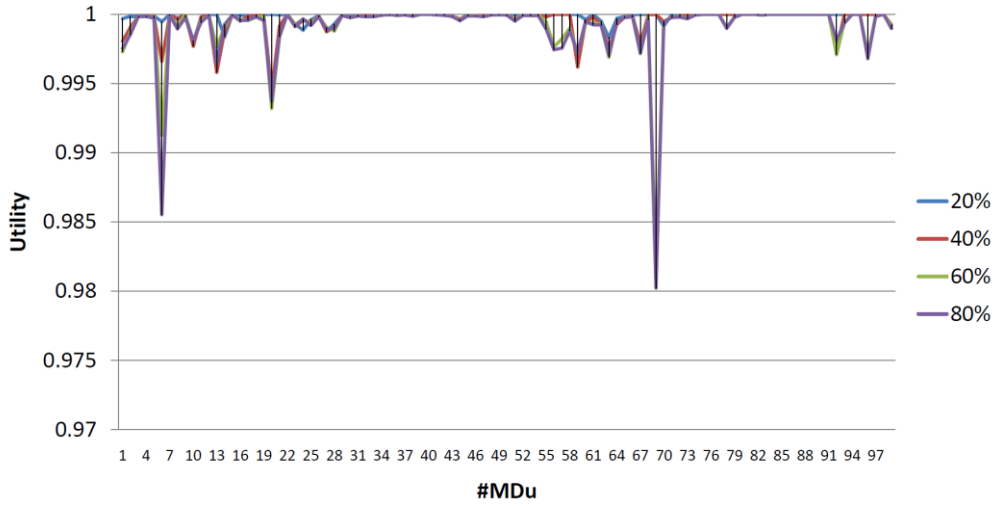


Figure 28: Utility evaluation.

As one can notice, the trade-off between privacy and utility is explicitly shown in the results where the increased percentage of anonymized words and multimedia objects decreases the utility of the multimedia documents. Nonetheless, such decrease of utility remains bounded by the number of words and multimedia objects contained in the multimedia documents signatures where only their content is subject to sanitization.

V.2.4 Evaluating Computational Cost

The MD^* -algorithm's time complexity is polynomial and of:

$$O(|\sigma_{E_u}(E)| \times (|W^*| + |MO^*| \times z)) \approx O(|\sigma_{E_u}(E)| \times (|W^*| + |MO^*|))$$

where:

- $|\sigma_{E_u}(E)|$ is the number of relevant multimedia documents retrieved from the external source,
- $|W^*| + |MO^*|$ is the number of sanitized words and multimedia objects from MD_u , and
- z is the number of attributes used by the selective intersections (which is limited and low).

This can also be seen experimentally in Figure 27. The resulting computational time depends on:

- 1) the conjunctive set of words and/or multimedia objects in E_u that are used to query the external source,
- 2) the external source from which multimedia documents (MD_β) are retrieved (e.g., the Web in our case). This is what we call fetching time which in some cases can be unpredictable as noticed between $\frac{1}{\alpha} = 4$ to 6 where the time to retrieve the individuals' data from the external source has increased (Figure 29).

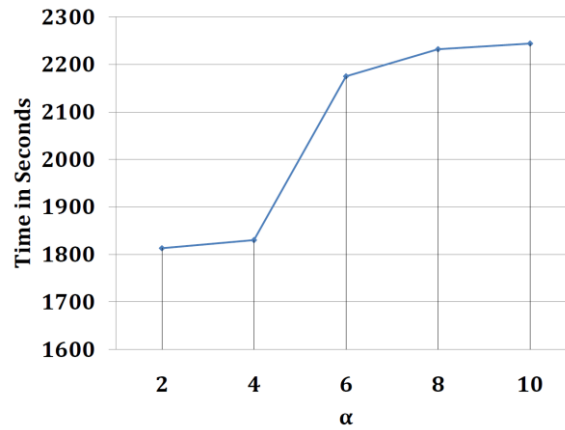


Figure 29: Computational cost evaluation.

VI. Conclusion

In this chapter, we proposed *de*-linkability, a privacy-preserving constraint that ensures the safe publication of multimedia documents. *de*-linkability addresses the privacy threat in its broader aspect while considering both textual and multimedia content.

We employed a selective intersection function to quantify the re-identification threat which is highly dependent on how much information can be acquired from i) adversaries' background knowledge and ii) external sources containing relevant information related to the anonymized individual.

In order to quantify common information between multimedia documents, we defined three operators:

- 1) to compare text content between documents such as “@fdbeauce” and *Beaucé* in the Wikipedia page about François Fillon,
- 2) to compare multimedia content according to multimedia objects such as the (geographical coordinates) metadata of the profile image of François Fillon and the metadata of the manor picture from Wikipedia page,
- 3) to match text and multimedia data such as “@fdbeauce” and the (geographical coordinates) metadata of the profile image of François Fillon.

Furthermore, we provided a sanitizing algorithm (*MD*-algorithm*) to protect against violating content and preserve at the same time a minimum quality through an adapted sanitization process that takes into consideration the complex nature of multimedia objects.

The evaluation of the efficiency of this algorithm was performed in terms of the uncertainty raised and the utility loss of multimedia documents due to the sanitizing process. With *de-linkability*, we were able to demonstrate that profile information of the Twitter users of the dataset tested cannot be used to infer the users’ identities. Additional information about an OSN user can be obtained by exploring events published and shared on his network. Therefore, the user profile can be updated or completed with information from events experienced by the user. In the next chapter, we will provide a methodology for detecting personal events in photos shared within online social networks, in order to enrich the profile information of the individuals to protect.

Chapter 4: Personal Event Detection from OSN Photo's Metadata

Abstract

Online social networking has become the predominant activity in the digital world thanks to multimedia data (mainly photos) sharing. Discovering events where users are involved using their own posts and those shared by their friends would be of great importance. In this chapter, we address this issue by providing an original approach able to detect user's events using photos shared within his online social networks. Using metadata, our approach provides a multidimensional gathering of similar photos using their temporal, geographical, and social facets. To validate our approach, we implemented a prototype called *Foto2Event* and conducted a set of experiments on the MediaEval dataset. Results show that our approach works well for various metadata distributions.

I. Introduction

Since photos are the most popular type of content shared on online social networks today, we propose in our study to exploit photos. Moreover, because photos uploaded online usually represent real stories from users' life, we exploit in this chapter personal events in the individuals' own posts as well as those shared by their friends/contacts. Specifically, we propose a clustering algorithm based on the photo collection's metadata. Our approach is able to detect and link users' *elementary events* using photos (and related metadata) shared within their online social networks. We particularly consider OSN photo's metadata that can be used to characterize *who* participated in the event, *where* the event happened, and *when* it happened.

There are several applications where events detected from photos may be useful, such as:

- **Enriching User profile:** User profile can be updated with information from events experienced by the user. For instance, it is common that many social network users do not complete their profiles with all personal information such as marital status and home location. However, information available about events (e.g., *event location*) can be used to disclose sensitive information for the user, or personal data he did not intend to disclose online (e.g., *home location*). This information, if found, would result in a richer user profile and thus prevent further the identity leakage in online social networks [142].
- **Detecting missing data:** Events may help in estimating missing (meta)data when photos do not have social tags or metadata [81, 143]. In essence, photos belonging to a particular event should have some identical information such as the location and date [144]. For instance, when a user is identified in a photo containing all metadata information, a propagation could be performed to estimate untagged faces and infer the date and the location of other photos detected in that same event.
- **Relationship discovery:** Some researches focused on identifying relationship types between users within the same social network or across different social network sites [145]. Yet, it is still a great challenge to discover "hidden" relationships between users

in a social network. For example, the Facebook algorithm presented in [146] uses the individual's network neighborhood to identify people who are dating. By finding people who usually attend the same events, uncovered relationships, as well as new relationships could be discovered.

In this study, we detect personal events in photos shared online to enrich the profile information of users. In essence, metadata of photos attached to events may include sensitive, private information about the user who needs to preserve his/her privacy. Therefore, exploiting users' events would provide better privacy protection to the anonymous users on social networks.

The rest of this chapter is organized as follows. In Section 2, we describe our data model used in our approach. A motivating scenario and key challenges are presented in Section 3. Section 4 gives a brief overview of our methodology. We present the pre-processing method in Section 5. Section 6 presents our event detection approach. Section 7 presents our prototype and experimental tests. Section 8 concludes this chapter.

II. Data Model

II.1 User (u)

u is a social network user who has an account on a social network site (e.g., Twitter, Facebook, etc.). Each user has an OSN user profile including personal information (e.g., *name*, *birth date*, *gender*, etc.), contact information (e.g., *email*, *mobile phone*, *address*, etc.), personal interests, work experiences, etc. (S)he can post content on the social networks in the form of photos, videos and text. Since we are interested here only in photos shared by users, we describe each user as follows:

$$u: (u_{Id}, pf, I)$$

where:

- u_{Id} : is the identifier of a given user. u can be identified on the social network site by his/her name, email address, or any other identifying information;
- pf : is the set of attributes used to describe the profile of a user and their corresponding values. This is defined in the Individual Profile definition (pf_u) in Chapter 3;

- I : is the collection of photos published or shared within his/her social network.

We note that this definition can be extended to include any type of multimedia data (e.g., video, audio, etc.).

II.2 Star Social Network (SSN)

We represent the social network of a given user u_0 as a star graph composed of the set of users directly connected to u_0 . In online social networks, people are connected to each other through links that can denote one or multiple relationships. These links can be of various types as in real-life (e.g., colleagues, relatives, friends, etc.). In this work, we assume that two users are linked by only one single type of relationship. Formally, we define the star social network of u_0 as follows:

$$SSN_{u_0}: (U, L, f_l)$$

where:

- U : is a set of users in the social network of u_0 ;
- L : is the set of labels describing the relationships between users;
- $f_l: U \rightarrow L$ is an association function mapping a link label to each person linked to u_0 .

II.3 Social Space (Δ)

The social space (Δ) is a coordinate system, representing the environment of a social network user u_0 , over three dimensions: locations, times, and person tags (Figure 30). It can be formally represented as follows:

$$\Delta: (\Delta_L, \Delta_T, \Delta_S)$$

where:

- Δ_L : is the set of locations identified in the social network of u_0 , ordered according to a given distance dist_{Δ_L} from u_0 's hometown. The units of Δ_L can be one of the following values: *GPS-coordinate*, *street*, *city*, *region*, or *country*. We assume here that dist_{Δ_L} values between two locations having the same distance from u_0 's hometown but within different directions are different. Different distances can be used such as the

Haversine's distance [125], the Euclidean distance between places [147], the Hausdorff distance between two finite set of points [148] , etc.

- Δ_T : is the set of the datetime values identified in the social network of u_0 , ordered according to a given distance dist_{Δ_T} from u_0 's date of birth. The units of Δ_T can be one of the following values: *hour*, *day*, *week*, *month*, or *year*. Exponential decay can be used to measure temporal distance [149, 150].
- Δ_S : is the set of users' names identified in the social network of u_0 , ordered by the alphabetical order. u_0 is the starting element. We assume here that each person has a unique name. A string-based distance dist_{Δ_S} can be used to decide whether the names are equal or disjoint.

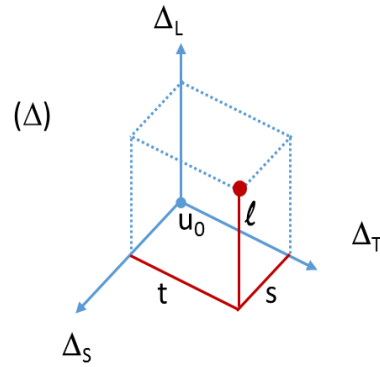


Figure 30: The social space Δ , modeled as a three-dimensional space: locations, times, and people names.

Δ satisfies the properties of a metric space (symmetry, non-negativity, and triangle inequality) since each dimension is associated to a distance (proof is omitted here since it is obvious).

II.4 Metadata (meta)

The metadata is a set of attributes describing a photo. Metadata contains context information about the creation of the photo (e.g., user identifier of the creator user, time and date of creation, place of creation, purpose of creation, description, technique, etc.). In this study, we consider only three of the 5W1H attributes: *who*, *where* and *when*, according to the social

space's features. Thus, we represent the metadata information of a photo as a point in the social space Δ :

$$meta: (l, t, s)$$

where:

- l represents the geo-location of a photo when captured (e.g., city, region, country, etc.);
- t represents the capture date/time of the photo. If the capture date/time is missing, we use the uploading date/time instead;
- s is the name of the creator or the publisher of the photo. In essence, the creator's name is used when (s)he belongs to the SSN_{u0} , otherwise the publisher name is used.

II.5 Image (img)

An image represents a photo posted on the social network of $u0$. It contains embedded metadata attributes that give information about the captured scene and people identified in it. In this study, we only focus on photos that indicate a place (i.e., l is not empty) or depict at least one person in SSN_{u0} within the captured scene. Formally, we represent an image as follows:

$$img: (imgId, meta, F)$$

where:

- $imgId$: is a unique identifier of the photo (e.g., its URI);
- $meta$: is the metadata describing the photo;
- F : is a set of people, belonging to SSN_{u0} , who are identified and tagged in the photo.

Figure 31 shows an example of a photo with its metadata.



Figure 31: Example of a photo with its metadata and faces of tagged persons.

II.6 Elementary Event (e)

We define the *elementary event* that will be used in this study as follows:

$$e: (ev_{id}, SB, meta_e)$$

where:

- ev_{id} : is the elementary event identifier;
- $SB(I', \Delta')$: is the story board describing the event e where:
 - $I' \subset I$: is the collection of photos taken in the event e and published on SSN_{u0} ;
 - $\Delta' \subset \Delta$: is the social space of the event e representing the geographical, temporal and social information related to this event such as:
 - $\Delta_L' \subset \Delta_L$: denotes the minimum bounding polygon (MBP) [151, 152] of the location in which the event took place. A method is disclosed in [152] for determining a MBP for use in a geo-coding system: a first polygon is received defining a geographic region. A minimal number of points is determined corresponding to the first polygon to define a second polygon at least bounding the first polygon.
 - $\Delta_T' \subset \Delta_T$: denotes the datetime interval in which the event took place. This interval is made from dates of the first and the last evidence (i.e.,

photo) that an event has taken place. Allen and Ferguson [153] presented a definition of temporal intervals, and defined events using interval temporal logic. In Allen's interval temporal logic, a time interval is defined in the linear time line, with a fixed starting point and ending point.

- $\Delta_S' \subseteq \Delta_S$: denotes the set of people participating in the event. We create bag-of-users for each event, by counting once each person tagged in the photos taken at the event.

Δ' can be visualized in a 3D graph (see Figure 32). The point set represents the photos' position corresponding to the same event.

- $meta_e$ is the set of representative moments, locations, and people/groups involved in the event. $meta_e$ is used to make inference information from the event social space. The inference information can be used in order to further identify the context of the event and thereby perform reasoning tasks which would aid the event semantization. By mining *significant-frequent* and *significant-rare* dates, locations and tags, it would be possible to identify interesting aspects and patterns which will provide a semantic description of events. Frequent and rare event aspects will be addressed in a dedicated study.

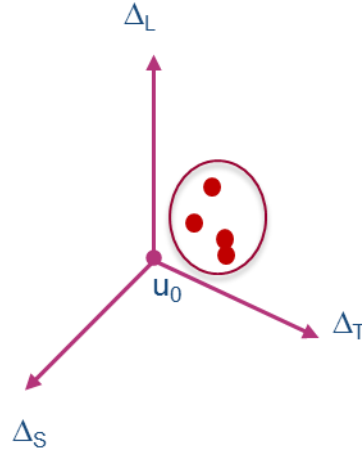


Figure 32: Representation of the event social space in a 3D graph.

III. Trip Scenario

Being an active user on Facebook, *Lisa* (u_0) frequently posts photos taken at different events. Her current residence is in the city of *Biarritz* in southern France. She has four hundreds of connections (e.g., family members, friends, colleagues, etc.). She posted photos taken with her colleagues and friends taken during a trip to *Sweden* in *May 2014*. Her colleagues are from *Biarritz* and her friends live in *Sweden*. To show the multi-* property of events, we use the Facebook photos of that trip. We illustrate here four major properties: i) multi-*source*, ii) multi-*participant*, iii) multi-*day*, and iv) multi-*site*.

III.1 The multi-* property of events

III.1.1 Multi-*source* Trip

Ghada, Kouki, Regi and Solomon are four colleagues who traveled with *Lisa* to Stockholm. They took photos and shared them through Facebook, but each in his/her own account. As a result, multiple albums of the same trip were created by different persons on SSN_{u_0} . We show some photos that illustrate this case in Figure 33.



Figure 33: Some photos shared on SSN_{u0} created by five different persons, including u_0 .

III.1.2 Multi-participant Trip

An event may include several participants; some of them post several photos about it while others do not. For instance, only 4 of the 12 colleagues who traveled with *Lisa* posted photos of the trip (as shown in Figure 33). The other colleagues did not post anything about it online. As we can see in Figure 34: *Saddem* is identified in the photos of *Lisa*, *Joseba* and *Irvin* are identified in the photos of *Regi*, etc.

However, it is important to note that *Lisa* does not appear in any photo with *Joseba* nor with *Irvin*. Since *Joseba* and *Irvin* are identified in the photos of *Regi*, and *Regi* has many photos with *Lisa*, one can conclude that *Joseba* and *Irvin* were involved in the trip event made by *Lisa*.

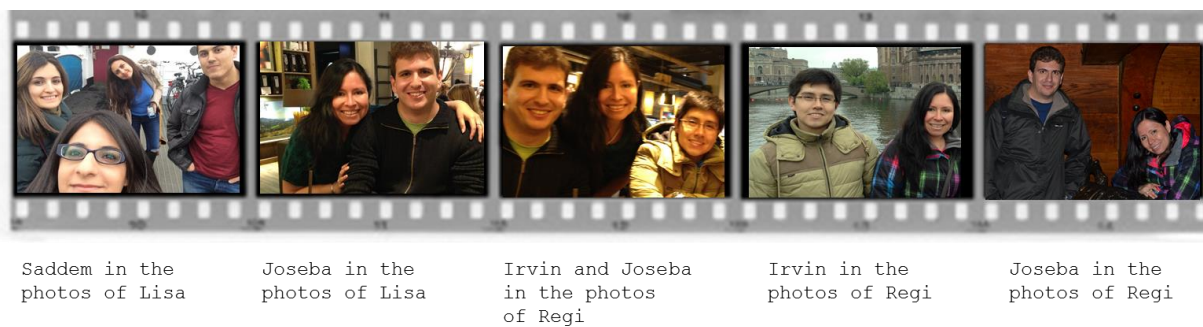


Figure 34: Some photos shared on SSN_{u0} created by *Lisa* and *Regi* where new participants (*Joseba* and *Irvin*) are identified.

III.1.3 Multi-day Trip

The trip to Sweden started on the 8th of May and ended on the 11th of May 2014. Lisa and her colleagues took photos on all days of the trip. Figure 35 shows photos of Lisa from the trip but with different creation dates.



Figure 35: Some photos shared on SSN_{u0} taken over the four days of the trip.

III.1.4 Multi-site Trip

The departure was from *Biarritz airport* in France to *Skavsta airport* in Sweden. Lisa took a photo with her colleague at *Biarritz airport* before take-off. When they arrived to Sweden, they took an airport coach that travels between *Skavsta* and *Stockholm*. Skavsta airport is located just outside of the town of *Nyköping*, about an hour and 20 minutes from downtown *Stockholm*. During her trip to Sweden, Lisa travelled to *Örebro* (in Sweden) to attend a wedding with old Swedish friends of her. Before the return trip to Biarritz, she took a group photo in *Skavsta airport* with her colleagues traveling with her. Consequently, Lisa has geotagged photos in five different cities shown in Figure 36, all taken during the same trip event.



Figure 36: Some photos shared on SSNu0 taken in different cities during the trip.

III.2 Challenges

This scenario shows that photos can include information about events that may be not available using only low-level features (e.g., colors, etc.) or the available description provided by the users. The main challenges that arise from this scenario are:

- Detecting photos depicting the same social events, given the metadata attributes describing photos and the person-tags. Given the large collection of photos shared in the social network of *Lisa*, the main challenge here is to identify all photos related to this trip, even though they concern *many participants*, or are published by *several sources*, and/or occur over *many days* and in *several places*. Considering this multi-* property ensures getting people, places and moments of the event that were not captured by the main user for many reasons (e.g., scenes subjectively not interesting for her, camera or cell phone battery running out of charge, weak social ties with some participants, etc.).
- Identifying all participants in the event:
 - whether they are identified in the same photos with u_0 or not:

This can be due to the type and the strength of their relationships with u_0 . In essence, the relationship strengths between users in online social networks –as

in real world– may vary significantly. This is the case of Joseba as described above.

- whether they posted about the event or not:

Events with many participants may include persons who are not active on social media or who do not share photos of their events online. This is the case of Irvin and Joseba as described above.

- Identifying the wedding event as a *sub-event* of the trip event to Sweden. This can be deduced from two facts:
 - The wedding took place in *Örebro* contained in Sweden during the same period of time as the trip event.
 - Friends of Lisa who are residing in Sweden posted photos of the wedding and tagged Lisa.

Sub-events and other related events will be presented in the next chapter.

IV. Methodology Overview

To overcome these challenges, we propose a clustering method to cluster photos based on their metadata and person-tags associated with them. Since metadata are most of the time automatically available in photos, we chose to rely on metadata to detect events from photos. Furthermore, metadata provide exact and objective information whereas annotations provided by users might be subjective (based on the user's feeling, interest, etc.). We also note that the processing cost (time and relevance) is low using metadata features.

An overview of our event detection algorithm is shown in Figure 37. The suggested algorithm is composed of two main steps: first, to define metadata and person-tags granularities; and second, to detect clusters of photos. The input is SSN_{u_0} , the social network of u_0 , for which we are interested in detecting events. The output contains linked elementary events of u_0 . In the following, we present our approach in more details.

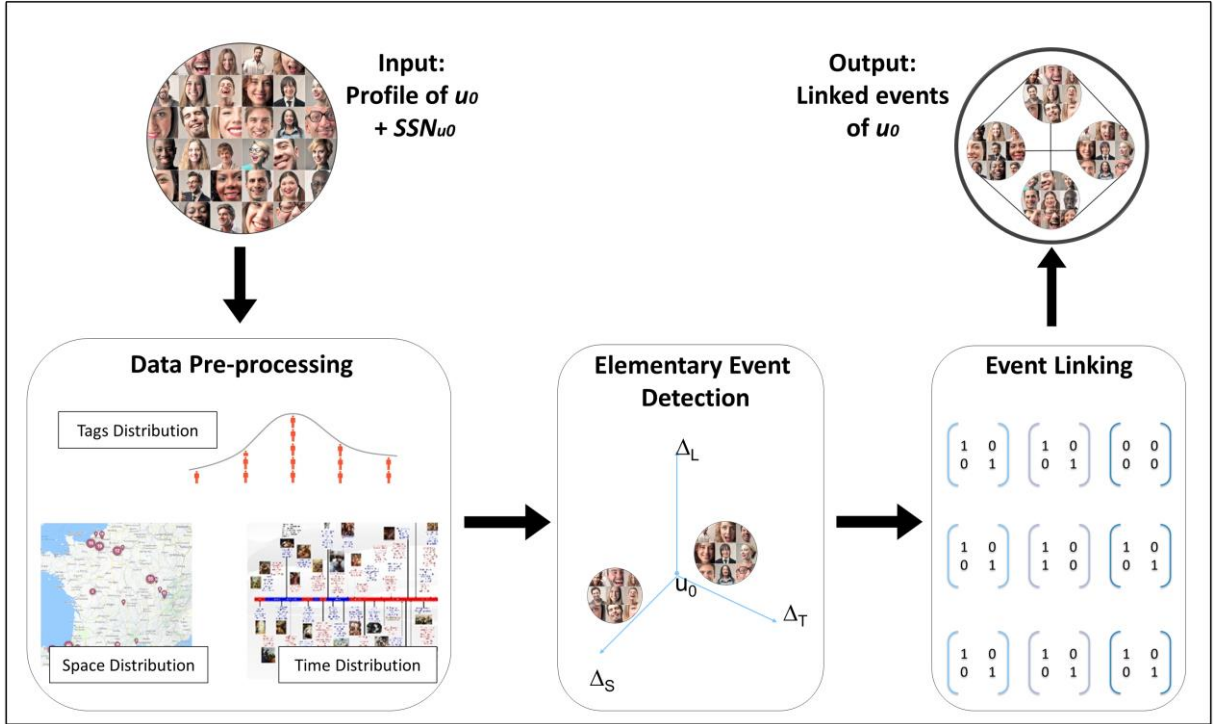


Figure 37: Description of our approach to detect elementary events.

V. Metadata Pre-processing

Identifying person-tag and metadata occurrences in the collection of photos would be useful to determine the appropriate parameters for the clustering. This pre-processing step takes into consideration the data distribution of photos we have at hand. Three factors can be analyzed: (i) the photo capture date/time, (ii) the photo locations and (iii) the people-tags depicted in photos.

V.1 Time Granularity

Firstly, we examine the dispersion of photos over time. The time granularity can be one of the unit values of the temporal axis Δ_T defined in the Social Space Δ , (such as a *year*, a *month*, a *day*, or an *hour*). If the time difference between photos corresponds to one day or more, we use *day* as the time granularity. Otherwise, we use finer time granularities such as *hour*.

V.2 Space Granularity

Secondly, we examine the geographic dispersion of the photos' locations. Locations can vary at different levels of granularity as the units used in the geolocation axis (Δ_L): *country*, *region*, *city* or *street*. Therefore, we also choose the space granularity based on the heterogeneity of photos' locations. If photos are taken in different streets, we use *street* as a granularity. Otherwise, we use finer space granularities such as *buildings*.

V.3 People-tags Granularity

Some people such as family or best friends are more likely to appear in the photos with u_0 than others. For this reason, the third purpose of the pre-processing algorithm is to identify frequent faces observed in the photos of u_0 . Therefore, we set a minimum value threshold $minF$. For each user u_i in the social network of u_0 , we measure the portion of photos where u_i appears in the collection. Frequent faces correspond to users who appear in the photos with u_0 in a portion p_i higher than $minF$. Other faces are considered rare faces. This measure will be more elaborated in a future work. In present work, only the time-space granularities are inputs to the clustering algorithm.

VI. Elementary Event Detection

After identifying the appropriate granularities based on the collection of photos at hand, this step aims at identifying clusters of photos belonging to the same event. In this section, we describe how to group similar photos together to form clusters corresponding to elementary events.

Different clustering methods are provided in the literature and can be divided into two broad categories: hierarchical and nonhierarchical. Hierarchical clustering algorithms produce nested sets of data (hierarchies), in which pairs of elements or clusters are successively linked until every element in the data set becomes connected [154, 155]. Nonhierarchical methods group a data set into a number of clusters irrespective of the route by which they are obtained. For instance, the k-means [156] groups the data set into k different clusters based on similarity, while density-based clustering [157] can create clusters with arbitrary shapes based on a metric space.

In this study, we propose an agglomerative, nonhierarchical clustering algorithm that uses three features based on our Social Space Δ : time feature, location feature, and creator/publisher feature. We define the time-space granularities as obtained from the pre-processing step. We use the Social Space corresponding distances between any two photos, $photo_i$ and $photo_j$: i) $dist_{\Delta T}$ between date/ time values, ii) $dist_{\Delta L}$ between locations, and iii) $dist_{\Delta S}$ between names. Our algorithm is based on the following assumption:

Assumption 1: Since a user can be only at one place at a time, photos captured by him/her at a specific time within a geographical boundary are considered as belonging to the same event.

Thus, a cluster k contains all photos created/published by a user on a common period and location. Formally:

$$\begin{aligned}
 k = \{photo\} / \forall i, j, \\
 dist_{\Delta T}(photo_i.meta_{\Delta}.t, photo_j.meta_{\Delta}.t) \leq \gamma \text{ and} \\
 dist_{\Delta L}(photo_i.meta_{\Delta}.l, photo_j.meta_{\Delta}.l) \leq \delta \text{ and} \\
 dist_{\Delta S}(photo_i.meta_{\Delta}.s, photo_j.meta_{\Delta}.s) = \varepsilon
 \end{aligned} \tag{8}$$

where:

- γ is a predetermined threshold related to the distance between the capture date/time values of two photos,
- δ is a predetermined threshold related to the distance between the geo-locations of two photos,
- ε is a predetermined threshold related to the distance between the creators' names of two photos. In this work, we set ε to 0.

VII. Prototype and Experimentations

In this section, we present the set of experiments we have performed to test the efficiency of our approach. First, we describe the prototype implemented to validate the clustering

algorithm. Then, we describe the dataset we used and introduce the evaluation strategy we followed. Finally, we present the experiments results.

VII.1 Prototype System

We developed a prototype system using Java, called *Foto2Event*, to test, evaluate and validate our event detection framework. In this section, we present the general architecture of our prototype (Figure 38), and we detail the different modules of the system: the *profile manager*, the *metadata manager*, the *photo parser*, the *preference manager*, and the *event builder*.

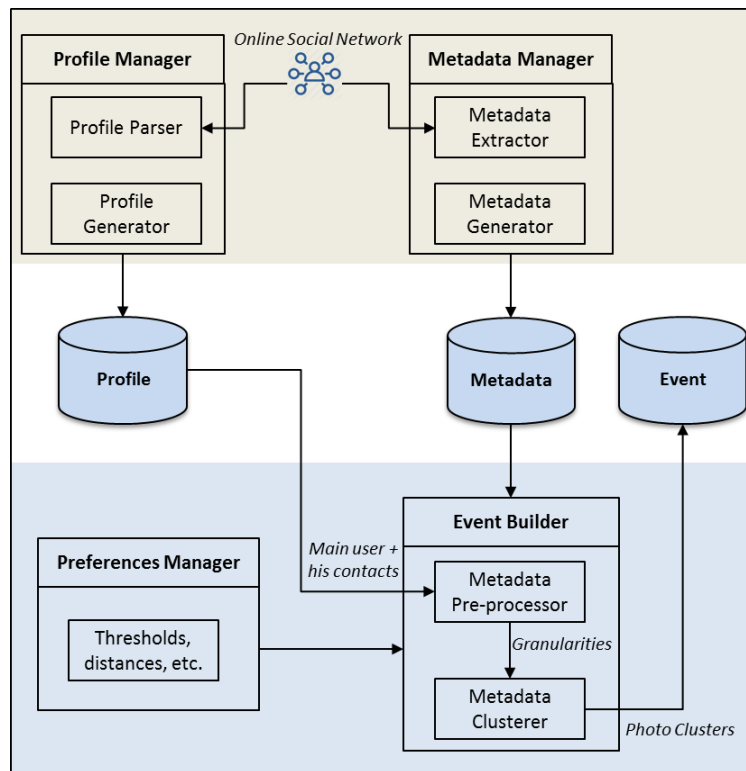


Figure 38: Architecture of our framework to detect events.

1. The *Profile Manager* component

The profile manager component manages profile information from one or several social networking websites. The role of this component is to: i) extract and store user information

from the social network profile, and ii) extract the contacts of the user in this network. This component allows also to generate from scratch or to complete a user profile within a star social network. As mentioned earlier, the user profile includes information such as *name*, *birthday*, *home address*, *interests* and *contact information*. The star social network includes a set of users (i.e. users' names) and their relations (e.g., friends, colleagues, etc.) with the main user.

2. The *Metadata Manager* component

This component allows extracting the photo metadata and stores them in XML files to provide input to the clustering algorithm. We chose XML because of its flexibility compared to relational databases. In addition, it is easy to parse and to validate (with XML Schema). This component can automatically generate custom metadata associated with photos such as creator name, location, and creation date, in various formats (e.g., geographic coordinates, cities, countries, etc.). Additional criteria may be applied to the generation of the photo metadata in order to highlight the behavior of our approach on different datasets (e.g., interval of n days for dates, small or large radius that delimitates the geographical neighborhood area for location, etc.).

3. The *Preferences Manager* component

Our system also has a Preference Manager that accepts some parameters of the algorithm such as the thresholds related to the distance between the capture dates or the geo-locations of photos, etc.

4. The *Event Builder* component

The event builder component accepts as input a photo metadata set. It first analyzes the distribution of the metadata to determine the corresponding granularities for each dataset. Then, it clusters photos' metadata based on their time stamps, geo-tags and creators to detect events. Each event is associated with a set of at least one photo.

VII.2 The MediaEval Dataset

The dataset we used is the *Social Event dEtection Dataset* (ReSEED) which was provided at MediaEval⁹¹ 2014. This dataset consists of a collection of 437,370 photos contributed by 4,926 unique Flickr users and assigned to 21,169 events in total.

We chose this dataset since it is a large real-world dataset of images, available online with their metadata for the purpose of detecting social events in large collections of multimedia items, and including labels for ground truth. The metadata includes the following: *username of the uploader, date, title, description, tags and geo-location*.

Before we get into the experimental evaluation, it is worthwhile to present some interesting and challenging aspects of the MediaEval dataset, as discussed in the ground truth statistics in [158]:

- Photos represent different types of events;
- The distribution of the events' duration is not uniform: Some technical events such as demonstration and protest events tend to last multiple days, while others like soccer events, last a few hours;
- The size of events varies significantly as well: While 3,598 events include only one single photo, and 1,799 events include 2 photos, there is a small number of events which include over 1,000 photos.

However, there is a significant difference between our case and the application of this dataset. The dataset contains pictures from the Flickr photo community site that is not a star network. In order to adapt the dataset to our case, we first chose one main user (u_0) from the 4,926 Flickr users. We chose the user who has uploaded more photos (1,497 photos). We parsed his photos and assumed that other users are his contacts in his Star Social Network. Relationships (or link labels) are assigned randomly. In future work, we plan to apply rules to derive relationship type between users.

⁹¹ <http://www.multimediaeval.org/>

One other missing information to complete in the MediaEval dataset is the tags of people in the photos. It is difficult to ask users to provide more information to obtain person tags identified in their photos. Therefore, it might be beneficial to search for this information in the *description* field provided for each photo. As it is a real-world dataset, the description is available for only a subset of the images (37.8%). For these reasons, we chose to leave for future work the interesting question of person-tags.

VII.3 Ground Truth Creation

Photos have metadata (creator name, latitude, longitude, date and time). Before conducting the experiments, we implemented a pre-processing algorithm in order to get a set of photos clustered into *elementary* events.

The input to this algorithm is a set of photos to be clustered into *elementary* events.

The first step is the creator segmentation. It aims at getting distinct creators from the image base. Therefore, we start by separating photos in segments based on their creator.

The second step is the spatial clustering. The goal of this step is to form clusters of photos which are near geographically. To do so, we use a non-supervised density-based spatial clustering using the DBSCAN algorithm. The parameters of the algorithm are: i) *epsilon*: this parameter specifies how close points should be to each other to be considered a part of a cluster. We fixed epsilon to 1 km. ii) *minPts*: this parameter specifies how many neighbors a point should have to be included into a cluster. In our case, we fixed *minPts* to 1. iii) *distance*: this parameter specifies the distance metric used. In our case, the distance used is the Euclidean distance.

The third step is the temporal clustering. This step corresponds to dividing clusters into subgroups which are close in time. Therefore, we calculate the time interval between all consecutive photos, the mean M and the standard deviation St of the time intervals. Then, we calculate the threshold t by summing the mean and the standard deviation of the time intervals between consecutive photos. Consecutive photos are separated into different clusters if the time interval is greater than a given threshold t .

This is described in the following pseudo-code fragment.

```

Input: Photo[], // a set of photos to be clustered into events
DBSCAN, Euclidean_distance, epsilon, minPts // clustering parameters

Output: Event[] // a set of photos clustered into events
epsilon  $\leftarrow$  1 // 1 Kilometer
minPts  $\leftarrow$  1
clustererInst  $\leftarrow$  CreateClustererInstance(DBSCAN, epsilon, minPts, Euclidean_distance)

C[]  $\leftarrow$  GetDistinctCreators(Photo[])

Sort(C[])
for each p  $\in$  C[] do
    I[]  $\leftarrow$  timestamp(pi+1) – timestamp(pi)

M  $\leftarrow$  Mean(I[])
St  $\leftarrow$  StandardDeviation(I[])
t  $\leftarrow$  M + St

for each c  $\in$  C[] do
    c.Segment[]  $\leftarrow$  getPhotos(c)
    c.Cluster[]  $\leftarrow$  CreateSpatialClusters(c.Segment[], clustererInst)
    for each cl  $\in$  c.Cluster[] do
        c.Event[]  $\leftarrow$  CreateTemporalClusters(cl, t)
return Event[]

```

VII.4 Evaluation Measurements

The main criteria used to evaluate the performance of our event detection algorithm are:

- the number of correctly detected events, and
- the number of correct/incorrect photos detected for these events.

Since the MediaEval dataset is accompanied by ground truth data, we compare the obtained sets of events to the ground truth sets of events. We use the two evaluation measures suggested in the MediaEval benchmark of 2014: *F-score* and *Normalized Mutual Information* (NMI). In the following, we provide the formula and discuss the role of each measure in details.

F-score is the harmonic mean of Precision and Recall for the retrieved photos. *Precision* (*PR*) identifies the number of correctly retrieved photos (but not the number of retrieved events), w.r.t. the total number of photos (correct and false) retrieved by the system.

Recall (R) underlines the number of correctly retrieved photos, w.r.t. the total number of correct photos, including those not retrieved by the system. Having:

- A the number of correctly retrieved photos (true positives),
- B the number of wrongly retrieved photos (false positives),
- C the number of correct photos not retrieved by the system (false negatives), *precision* and *recall* are computed as follows:

$$PR = \frac{A}{A+B} \in [0,1] \quad R = \frac{A}{A+C} \in [0,1] \quad F - value = \frac{2 \times PR \times R}{PR+R} \in [0,1]$$

High *precision* denotes that the clustering algorithm achieved high accuracy in retrieving correct photos, whereas high *recall* means that very few correct photos were missed by the system. In addition to evaluating *precision* and *recall* separately, it is a common practice to consider *F-score* as a combined measure, representing the harmonic mean of *precision* and *recall*. High *precision* and *recall*, and thus high *F-value* indicates in our case high clustering quality.

Another performance measure used to evaluate the clustering accuracy is the **Normalized Mutual Information (NMI)**. In information theory, mutual information is a measure of the information overlap between two random variables [159]. The MI between random variables X and Y , whose values have marginal probabilities $p(x)$ and $p(y)$, and joint probabilities $p(x, y)$, is defined as:

$$MI(X; Y) = \sum_{x,y} P(x, y) \ln \frac{p(x, y)}{p(x) p(y)} \in [0,1]$$

This measure score is also in the range $[0, 1]$. Higher values indicate a better agreement with the ground truth results. This is because the mutual information between two sets of clusters becomes larger as one set of clusters is more consistent with the other set.

Recent studies [160, 161] have shown that the mutual information measure, being sensitive to the amount of overlap between two sets, can be normalized. Normalized mutual information (NMI) has been widely used in a lot of applications to measure the performance of clustering methods [162-164]. In our experiments, we apply the NMI measure to check the overlap between our clustering result and ground truth clusters.

A good clustering should have high recall and precision, and also high similarities with the ground truth. F-score measures only the goodness of the retrieved photos but not the number of retrieved events, nor how accurate the correspondence between retrieved photos and events is. Whereas, NMI considers the goodness of the retrieved photos and their assignment to different events [165]. Hence, we use both measures to measure the performance of our algorithm and to compare it with existing algorithms.

VII.5 Experimental Evaluation

In this section, we present the set of experiments conducted to evaluate the efficiency of our approach. The goals of these experiments were:

- 1) To measure the clustering quality for different cases considering i) the metadata distribution, and ii) the metadata availability,
- 2) To determine the time performance of our event detection algorithm, and
- 3) To compare our approach with other event detection approaches.

VII.5.1 Event Clusters' Quality

In order to measure our clustering quality, we run first a series of experiments to test our algorithm on different data distributions. Then, we run another series of tests to evaluate the quality of the clustering considering the availability of metadata items used in our approach.

VII.5.1.1 Experiment 1: Testing on different data distributions

The aim of this test was to study the effect of metadata distribution on detecting events using our approach. To test this, we generated three groups of data from the MediaEval dataset.

Each group satisfies specific criteria based on the distribution of the time, location or uploader information. The characteristics of these datasets are summarized in Table 7.

Table 7: Distribution of metadata in each dataset group.

Group	Variable	Values
Group 1	Number of distinct years	1 year (2006)
		2 years (2006, 2007)
		3 years (2006, 2007, 2008)
		4 years (2006, 2007, 2008, 2009)
		5 years (2006, 2007, 2008, 2009, 2010)
		6 years (2006, 2007, 2008, 2009, 2010, 2011)
Group 2	Number of distinct countries	1 country (Australia)
		2 countries (Australia, US)
		3 countries (Australia, US, Ireland)
		4 countries (Australia, US, Ireland, Spain)
		5 (Australia, US, Ireland, Spain, Sweden)
Group 3	Number of distinct creators	100 users
		200 users
		300 users
		400 users
		500 users
		600 users
		700 users
		800 users
		900 users
		1000 users

Group 1: The distribution over time is not constant in the original dataset. It consists of images from Flickr with an upload time between January 2006 and December 2012. Therefore, we created six groups of datasets, where each dataset consists of a set of images

uploaded in one or several years. We tuned the number of distinct years in order to collect photos taken during the same year(s). We fixed the space granularity to *city*. Then we ran the algorithm using *year* as the time granularity, followed by *month*, then *day*. The results are shown in the graphs in Figure 39 and Figure 40.

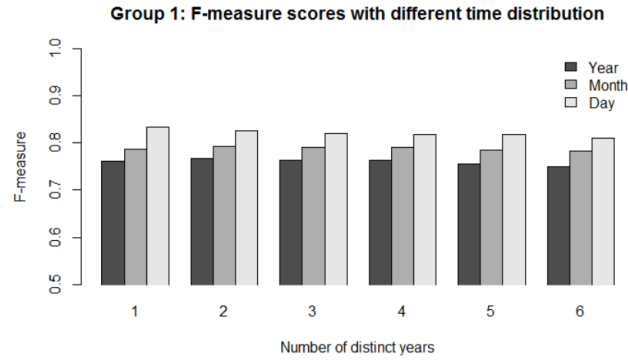


Figure 39: F-measure scores considering different time distributions.

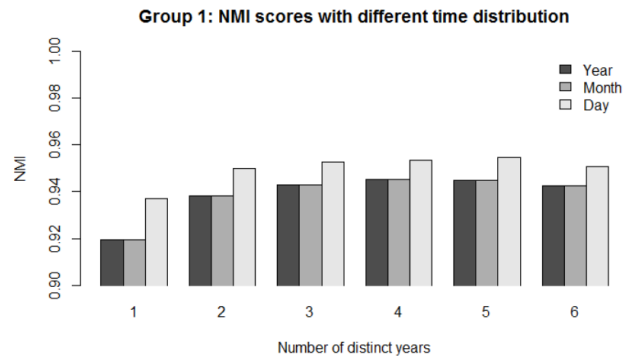


Figure 40: NMI scores considering different time distributions.

Results in Figure 39 show that f-measure ranges between 0.75 and 0.77 with the *year* granularity. It increases when using the *month* granularity and ranges between 0.78 and 0.79. However, our method yields the highest value (0.83) with the *day* granularity. Similarly, results in Figure 40 show that NMI ranges between 0.92 and 0.94 with the *year* granularity. It increases when using the *month* granularity and ranges between 0.93 and 0.95. Our method

also yields the highest NMI value (0.95) with the *day* granularity. Hence, the best performing result with different time distributions is obtained when using the *day* granularity.

Group 2: Photos in the dataset are heterogeneous and their location information may be significantly different. For instance, the location data may be over-close (e.g., when almost photos are taken in one city) or too scattered (e.g., when photos are taken in various countries around the world). Therefore, we created five sub-datasets, where each sub-dataset consists of a set of images captured in one or several countries. We tuned the number of distinct countries in order to collect photos captured in the same countries. We fixed the time granularity to *day* based on the previous result (Experiment 1). Then we ran the algorithm using *country* as the space granularity, followed by *city*, then *street*. The results are shown in the graphs in Figure 41 and Figure 42.

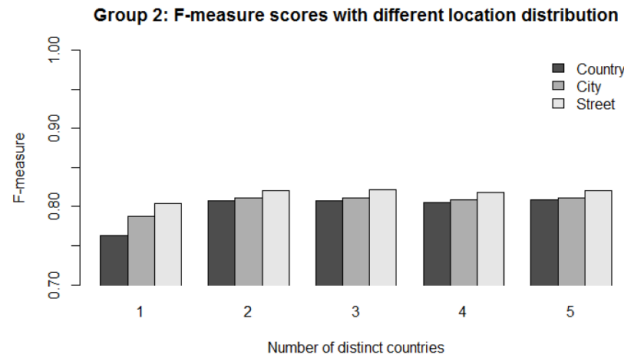


Figure 41: F-measure scores considering different location distributions.

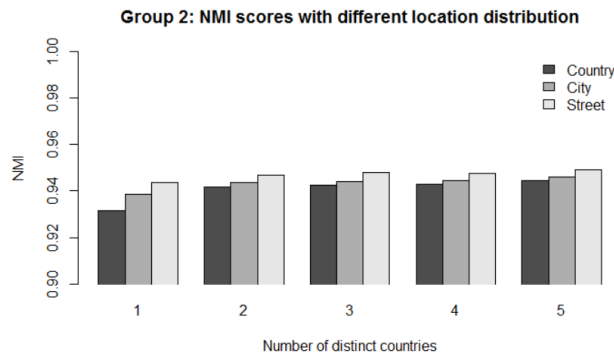


Figure 42: NMI scores considering different location distributions.

Results in Figure 41 show that f-measure values range between 0.76 and 0.81 when using the *country* granularity. We observe that our method yields better f-measure values with the *city* granularity and ranges between 0.79 and 0.81. However, when using the *street* granularity, f-measure is 0.82 for almost all tests. Similarly, our method yields the lowest NMI (0.93) when using the *country* granularity and almost the same values (0.94 and 0.95) with the *city* and the *street* granularities as shown in Figure 42. Hence, we observe the best f-measure and NMI scores achieved for almost all tests when using the *street* granularity.

Group 3: As already mentioned above, the photos in the dataset were uploaded by different users on Flickr. First, we collected the photos of the main user – the one who has uploaded more photos. Then, we included photos which are uploaded by a specific number of other users chosen randomly. The number of distinct users, including the main user varies from 1 to 1000. Following the previous results, we fixed the time-space granularities to *day* and *street* (based on Experiment 1 and Experiment 2). Then, we ran the algorithm. The results are shown in Figure 43.

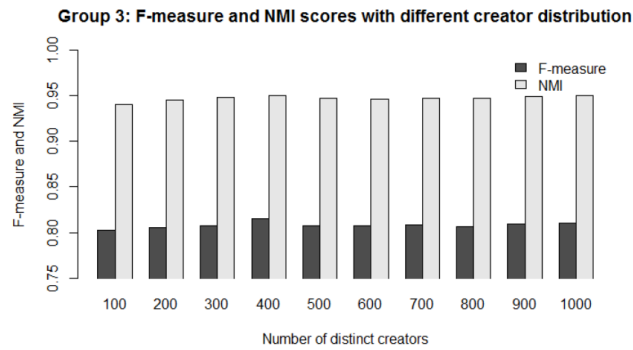


Figure 43: F-measure and NMI scores considering different uploaders distributions.

The graph in Figure 43 shows that NMI and f-measure have nearly the same value for all tests (0.8 and 0.95 respectively). This is because results depends on the values of the “target” variables (time, location, and creator) rather than the number of distinct uploaders.

VII.5.1.2 Experiment 2: Considering metadata availability

The aim of this test was to evaluate the benefit of having the metadata available in photos (creator, time and geographic information). We used the MediaEval dataset described previously. The collection of images was subject to the clustering algorithm based on their metadata. As it is a real-world dataset, there are some metadata that are not present in all images. For instance, the capture time is present for 98.3% of the photos, while the uploader information (i.e., username) is available for every photo. However, the location information is present in only 45.9% of the photos in the collection. We generated 5 groups of datasets based on the availability of metadata. The characteristics of these datasets are summarized in Table 8.

Table 8: Availability of metadata in each dataset.

Groups	Datasets	Capture time (T) %	Geographic information (G) %	Uploader information (S) %
Setfull	SetFull T, G, S=100%	100	100	100
SetFull (T)	SetFull T=75%	75	100	100
	SetFull T=50%	50	100	100
	SetFull T=25%	25	100	100
SetFull (G)	SetFull G=75%	100	75	100
	SetFull G=50%	100	50	100
	SetFull G=25%	100	25	100
SetFull (S)	SetFull S=75%	100	100	75
	SetFull S=50%	100	100	50
	SetFull S=25%	100	100	25
SetFull (T, G, S)	SetFull T, G, S=75%	75	75	75
	SetFull T, G, S=50%	50	50	50
	SetFull T, G, S=25%	25	25	25

Setfull dataset: The first group consists of the set of photos that are geotagged with location information (consisting of a pair of latitude-longitude coordinates), uploaded by a specific Flickr user, and provided with a (capture) time information. The objective of this experiment is to show the best case behavior when all photos' metadata are available for the event detection algorithm. Once again we tested our algorithm on all the possible granularities of time and space to confirm our choice of granularities (*day* and *street*). A qualitative comparison is shown in Table 9. Because in our experiments, we consider three time granularities (i.e., *day*, *month*, *year*) and three space granularities (i.e., *street*, *city*, *country*), we performed 9 tests. For each test, we used different combination of the granularities (the 2nd and 3rd columns), and measure the f-measure and NMI values.

Table 9: F-value and NMI scores considering different granularities of time and space on the Setfull dataset.

Test #	Time granularity	Space granularity	F-measure	NMI
1	day	street	0.8129	0.9519
2	day	city	0.8101	0.9514
3	day	country	0.8125	0.9525
4	month	street	0.7923	0.9498
5	month	city	0.7818	0.9482
6	month	country	0.7532	0.943
7	year	street	0.7923	0.9498
8	year	city	0.7485	0.9426
9	year	country	0.6112	0.9136

The experimental results showed reliable results for different granularities when there is no missing metadata. Specifically, we obtained the best results of f-measure (0.8129) and NMI (0.9519) using the *day* and *street* granularities. This experiment confirms the results of the previous experiments by again demonstrating that our method with the *day* and *street* granularity produces the best quality of clustering. For this reason, in what follows we test the clustering algorithm using only the *day* and *street* granularities.

SetFull (T) dataset: The second dataset group consists of a set of photos which are geotagged with location information and uploaded by a specific Flickr user. However, the capture time information is present in 75%, 50% and 25% of the photos (SetFull T=75%, SetFull T=50%, and SetFull T=25% respectively). We modified the Setfull dataset to meet these percentages. This group was created to test the usefulness of the capture time information for the event detection. The results are given in Figure 44. The f-measure and NMI values appear to be sensitive to the availability of the time information: f-measure decreases from 0.81 (100%) to 0.74 (75%), then from 0.61 (50%) to 0.44 (25%). Similarly, NMI value decreases from 0.95 (100%) to 0.91 (75%), then from 0.88 (50%) to 0.85 (25%).

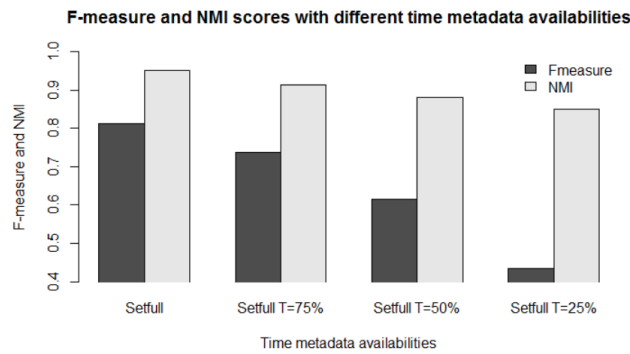


Figure 44: F-measure and NMI results considering different temporal information availabilities.

SetFull (G) dataset: The third dataset group consists of a set of photos which are uploaded by a specific Flickr user, and provided with capture time information. However, the geographic information is present in 75%, 50% and 25% of the photos (SetFull G=75%, SetFull G=50%, and SetFull G=25% respectively). Similarly, we modified the Setfull dataset to meet these percentages. This group was created to test the usefulness of the geographic information for the event detection. The results are given in Figure 45. Similarly to the time availability, the f-value and NMI levels decrease with the availability of geographic information: f-value decreases from 0.81 (100%) to 0.74 (75%), then from 0.61 (50%) to 0.44 (25%). Similarly, NMI value decreases from 0.95 (100%) to 0.91 (75%), then from 0.88 (50%) to 0.85 (25%).

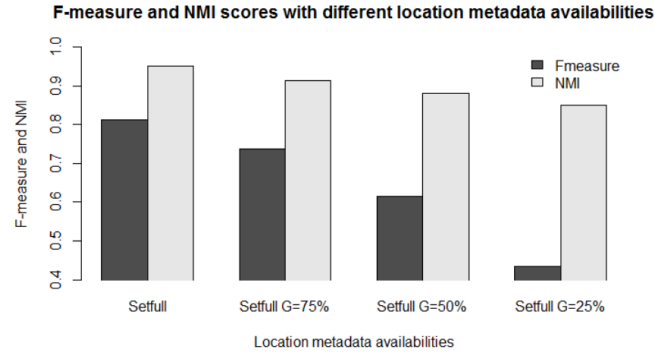


Figure 45: F-measure and NMI results considering different geographic information availabilities.

SetFull (S) dataset: The fourth dataset group generated consists of a set of photos which are geotagged with location information and provided with capture time information. However, the uploader information is present in 75%, 50% and 25% of the photos in the datasets (SetFull S=75%, SetFull S=50%, and SetFull S=25% respectively). Similarly, we modified the Setfull dataset to meet these percentages. This group was created to test the usefulness of the uploader information for the event detection. The results are given in Figure 46. The same behavior can be observed in the case of the uploader information: f-measure decreases from 0.81 (100%) to 0.74 (75%), then from 0.61 (50%) to 0.44 (25%). Similarly, NMI value decreases from 0.95 (100%) to 0.91 (75%), then from 0.88 (50%) to 0.85 (25%).

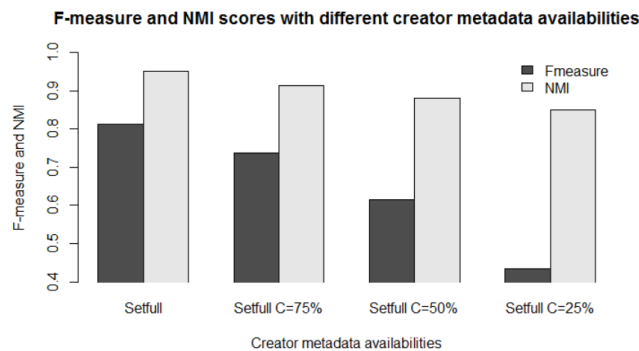


Figure 46: F-measure and NMI results considering different uploader information availabilities.

SetFull (T, G, S) dataset: The fifth dataset group consists of the three worst cases where the time information, the location information and the uploader information are provided in 75% of the photos, in 50% of the photos, and then in 25% of the photos only (SetFull T, G, S=75%, SetFull T, G, S=50%, and SetFull T, G, S=25% respectively). The objective of this experiment is to show the worst case behavior when some photos' metadata are missing for the event detection algorithm. The results are given in Figure 47.

F-measure levels decrease from 0.81 to 0.74 to 0.61 then to 0.44 where the location, time and uploader information is available in 100%, 75%, 50% then 25% of the photos respectively. Similarly, NMI levels decrease from 0.95 to 0.91 to 0.88 then 0.85 respectively. The above results emphasize the usefulness of having at least two of the metadata, in order to optimize the clustering process.

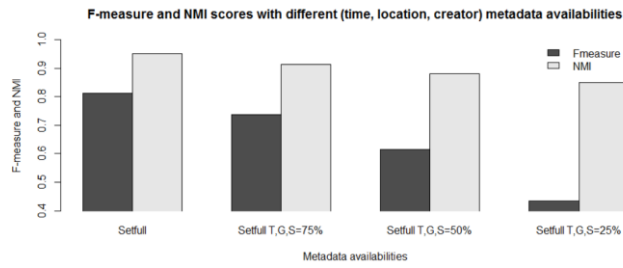


Figure 47: F-measure and NMI results with different metadata availabilities.

VII.5.2 Time Analysis

We implemented the event detection algorithm code in Java and conducted experiments using Intel Core i5 CPU @2.4 GHz with 8 GB RAM.

We experimentally analyzed the execution time complexity of our event detection algorithm. We varied the number of photos in the collection.

In order to evaluate time performance of the system w.r.t. the size of the photo collection, we extracted 10 random samples of the Setfull dataset (photos with metadata). We applied our

algorithm 10 times on each sample. Then we measured the average execution time required by our system, including both CPU and I/O time in order to verify our approach's linear time dependency on the size of the photo collection. Figure 48 shows that the time needed to perform clustering of photos in a dataset grows in a linear fashion with the dataset size.

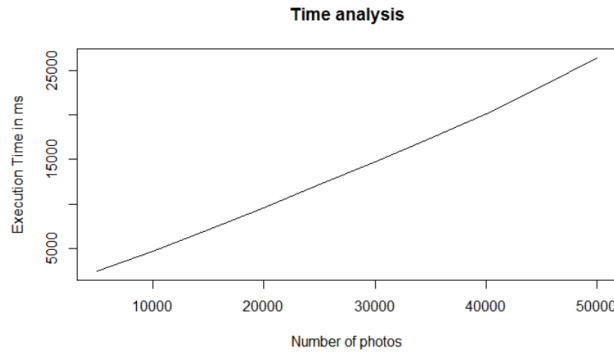


Figure 48: Time performance w.r.t. dataset size.

VIII. Conclusion

This chapter provides a clustering algorithm for automatically detecting events using photos shared online on the social network of a specific user.

The Social Space is introduced, with other definitions necessary to present our approach, such as star social network, photo's metadata, etc. Using metadata, our approach provides a multi-dimensional clustering of similar photos based on their temporal, geographical, and social facets. Firstly, we presented a pre-processing algorithm to determine the appropriate time-space granularities. This allows detecting clusters at a granularity that makes sense to the geographical and temporal distribution of the user's image data. Secondly, we showed how we group photos together based on the three features, related to our Social Space. At this stage, we obtained *elementary events*. More semantic-meaningful clusters can be obtained by building relations between clusters of events. This is what we will propose in the following chapter.

Chapter 5: Linking Elementary Events

Abstract

Events are increasingly attracting researchers today. Many studies and benchmarks were conducted to detect events from posts people are submitting online at any occasion they have. However, these approaches do not exploit the links that can exist between events. Links can be related to any aspect of events. Computing links between events on online social networks can help collecting more personal information and enforcing users' privacy. In this chapter, we introduce our methodology for describing links between events. First, we describe our meta-model based on the 4-Intersection Model. Then, we present our methodology to identify basic relations based on three main aspects of events: spatial, temporal and social. We will show how we can combine those basic relations to infer more complex relations such as ontological relations and application-based relations.

I. Introduction

As mentioned in the Event Definitions section in the Related Works chapter, an event is a symbolic abstraction for the semantic segmentation of happenings in a specific *spatio-temporal* volume of the real world [178]. It includes the presence of entities (i.e. *persons* who attended or witnessed the same event) throughout the same time and space. Today, people can share photos about their events with many different Web 2.0 tools, such as Facebook, Instagram, and Flickr. However, there is no tool that allows to link events together based on general characteristics or aspects of events (e.g., events occurring at a particular time), using classical ontological relations (e.g., link events by cause and effect), or via user-defined semantics (e.g., link birthday events). Event detection requires examining events and their inter-relations as well. For this reason, we present in this chapter a meta-model that allows representing, combining and inferring inter-event relations in an expressive and flexible way. We aim to discover relations that can exist between events detected in photos shared online. Thanks to our meta-model, we can define three types of event relations: i) basic relations, ii) ontological relations, and iii) application-based relations.

We propose an approach to automatically generate relations that correspond to different aspects of elementary events based on a homogeneous representation. In this work, we focus on the three aspects of elementary events: spatial, temporal and social. In addition, we present a methodology that can combine spatial, temporal and social relations of the event for modeling complex relations (e.g., link sub events that are separated in space but are part of the same event). To reach that goal, we identify a list of ontological relations, such as *caused-by*, *is-a*, etc. We model relations with a binary representation, which is supposed to be easy, flexible and extensible. Therefore, our approach can be applied for identifying arbitrary relations that can be defined by the social network user based on her context, his/her applications or his/her preferences as well.

Computing links between events is useful through various Web applications and can be used to support several application needs:

- **Collecting personal information from online social networks:**

Many studies [179-181] have tried to extract personal information from social media websites to provide better profiling (e.g., tag-based user profile, content-based user profile, etc.). However, the ultimate challenge remains: how to get richer user profiles to estimate missing information? By examining relations between events, we can infer unknown characteristics or traits about a user based on event information in order to construct or complete a user profile. For example, when the user's events always occur in the same neighborhood, we can classify the person as a "stayer". Otherwise, we define the person as a "traveler". Another example is when a user often goes to watch drama movies (as events), we can automatically update his/her preference to this kind of movies.

- **Enforcing privacy in online social networks:**

Social network users might make careful choices about whom they disclose photos about their events to. These choices might be dependent on the nature and the strength of the relationship they have with their contacts (e.g., friends, close friends, acquaintance, etc.). By computing links between events, social web platforms will be able to offer the user better custom privacy settings that allow each contact to only view events that they experienced with the group they belong to. For example, the user can choose to share photos of events occurring at work only with his/her colleagues. Therefore, his/her colleagues will not be able to view his/her family-related events by using this kind of privacy setting.

- **Understanding user behaviors in online social networks:**

Novel methods [182] were used to perform social network analysis (SNA) to understand how users are influenced online. These methods usually provide quantitative measures [183] based on statistical analysis on the available data, and do not focus on the visual representation of data. There are also structural metrics that are applied based on mathematical properties of the social network (e.g., centrality measures, structural groups, etc.). In a recent Facebook experiment in July of 2015,

over a million people changed their Facebook profile pictures to a rainbow filter to support the *American Supreme Court decision event legalizing gay marriage*⁹². By setting up this tool, Facebook was able to track users support for this event and thereby got an unprecedented insight on how to influence their users, as a kind of a psychological testing. This example shows how events detected in photos shared on online social networks can be used to understand their users' behaviors.

- **Expressive querying among a large number of photos:**

Adding links between events is useful to users who are interested in looking at their photos based on specific criteria such as time, place, or people involved. As the number of photos uploaded per user is becoming increasingly huge, it becomes more difficult to manage and search for particular events. Therefore, establishing links between events would allow the user to answer queries quickly and effectively from large amounts of data. These queries can be, but not limited to the following:

- a) Which events are *co-located* in the same place with one specific event?
- b) Which events are *co-occurring* in the same time period as a given event?
- c) Which events are *experienced by the same users* as the ones participating in a given event?

By linking events, we can answer those queries and also solve the multi-* property of events, which is a major problem in the event detection as illustrated in Chapter 4.

In spite of wide range of research on the event detection on online social networks shown in the Related Works chapter, only a limited number of studies were conducted on the event relations. Currently, no one has proposed how event relations can be incorporated into the Social Web.

⁹² <http://conservativepost.com/everyone-who-changed-their-facebook-photos-to-rainbow-just-got-duped/>

The rest of this chapter is organized as follows. In Section 2, we present definitions that are used throughout the chapter. In Section 3, we describe how we developed our meta-model. Then, in Section 4, we present basic relations between events and how to identify them. In Section 5, we present a methodology to combine relations according to some constraints that we have formalized. We show how we translate this to a binary representation in Section 6. In Section 7, we suggest a logical expression to represent ontological relations and define some application-based relations that can exist between events. In Section 8, we conclude and summarize the benefits and extensions of this meta-model and approach.

II. Definitions

II.1 Event Context: Who, Where, When

Events in social media can be divided into two types: i) explicitly created events and ii) implicit events [184]. In the first type, people create event pages (on Facebook, LinkedIn, Google+, etc.) to drive attendance and provide event details (schedules and information), or to promote events before, during and after they take place. Our focus is on events of the second type, where social network users add content (such as photos) during or after attending the event, without explicitly creating the event. In this case, it is important to understand the context in which these events take place. This understanding of the context could serve to identify, among others, the relations between events and to answer many questions like: *What other events is this event related to and how?* While many definitions of context have been proposed in the literature [185], most include three elements: location, time and people. This corresponds to the previously defined social space Δ in Chapter 4.

In the following, we introduce our relational social space that we will use in this chapter.

II.2 Social-R Space

Let R denote a set of all possible distinct relations that might exist between two events in a star social network. Each relation $r: \langle name, v_{\Delta R} \rangle$ is defined by its *name* and its value $v_{\Delta R}$. $v_{\Delta R}$

is a three-dimensional value (l, t, s) where l , t and s are values defined in the Social-R space Δ^R defined as follows.

A Social-R Space (Δ^R) is a coordinate system for a three-dimensional space representing the Socio-Spatio-Temporal relations between two events in a social network such as:

- Δ_L^R : represents the spatial relations that exist between two events;
- Δ_T^R : represents the temporal relations that exist between two events;
- Δ_S^R : represents the social relations that exist between two events.

Each point $r(l, t, s)$ in the three-dimensional space represents a particular combination of those three relations. The scales of the three axes are the same as shown in Figure 49. l , t and s are hexadecimal values and $l, t, s \leq F$. The choice of this scale will be discussed later.

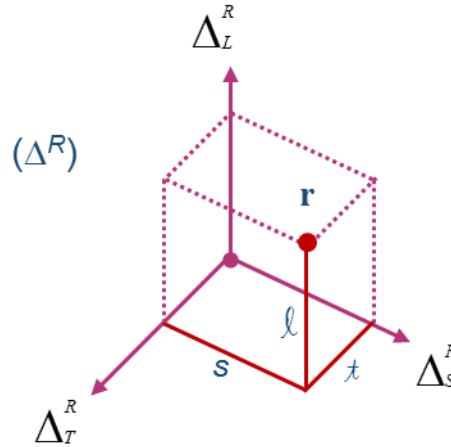


Figure 49: Representation of the social-R space in a 3D graph.

Using this three-dimensional space, a set of event relations can be generated within it based on the 3Ws: *Who*, *Where*, *when*. It is worthy to note that the present method can be extended to n -dimensional data if we have more contextual information about events (e.g., *What*, *Why*, *How*, etc.).

Consequently, we can easily move from the first space Δ to the second space Δ^R in a homogeneous method (see Figure 50). The main interest of this representation is to express clearly exclusive⁹³ relations that exist between events using their metadata, and thus to move from *image raw data* to *Linked Event Data*.

The advantages of our model are as follows:

- (A1) A simple and not expensive hexadecimal representation;
- (A2) An easy-to-read and expressive fashion;
- (A3) An extensible model for more complex events relations.

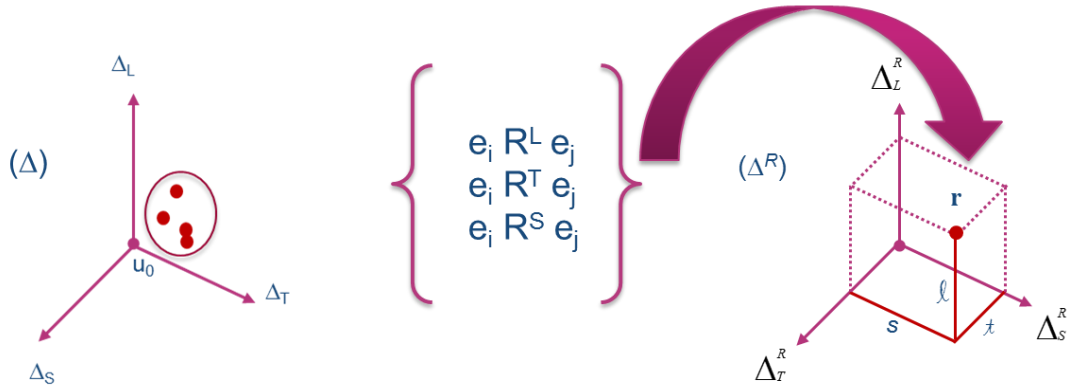


Figure 50: Transition from the social space (on the left) to the social-R space (on the right).

III. Meta-model of Event Relations

We present a meta-model to derive relations between events. Our meta-model is based on the 4-intersection formalism for topological relations to represent temporal, spatial and social relations between events. In the following, we emphasize on the importance of having a meta-

⁹³ There is one relation that links two events.

model description of event relations. Then, we introduce the neighboring sets and the 4-intersection model for binary topological relations.

III.1 Why a Meta-model?

We propose a unique meta-model as a basis for representing relations between events. The meta-model allows to deal with many different aspects of events (temporal, spatial, social, causal, etc.) in similar way. Moreover, it can support textual (e.g., author, description, etc.) and non-textual features (e.g., creation time, geographical coordinates, etc.). Each feature is associated with its appropriate distance metrics using different distance thresholds, at different levels of granularity. For instance, a pair of features are *disjoint* if the distance between them is less than a threshold for some distance metric (e.g., the number of minutes elapsed between two date/time values, the Haversine distance between two locations, etc.).

By using a meta-model based approach, our proposal is extensible for more heterogeneous data sets and additional contextual parameters. Thus, it allows us to comprehend events and express links among them in a robust and scalable fashion. Thus, the purpose of a unique meta-model is motivated by two primary needs:

- 1) Properly represent the heterogeneous aspects of events, and
- 2) Represent various relations in a unified and flexible way.

III.2 Topological Models

Several topological models [186, 187] have been developed, from which the most used are the 4-intersection model and its extensions, such as the 9-intersection model [188].

- The 4-Intersection model (4IM) represents topological relations by calculating the four intersections of the interior and the boundary of two objects (i.e., emptiness and non-emptiness). This initial model for binary topological relations was developed for 2-dimensional objects (e.g., two regions, etc.). It consists on intersection sets and an intersection vector.

- The 9-intersection model (9IM) is an extended representation of the 4-intersection model. It adds to the 4-Intersection model additional intersections related to the exteriors, where the exterior is obtained relative to the embedding space. This model was then introduced to enable the identification of more detailed relations when one or both objects are embedded in higher-dimensional spaces (e.g., a line and a real object). Relations are described here by the nine set intersections (intersections of the interiors, boundaries, and exteriors), instead of the four set intersections (intersections of the interiors and boundaries only).

In this work, we will not deal with exteriors at all since we are not interested in relations with respect to the embedding space. Furthermore, the 4IM provides a high degree of expressivity in our case, with a small cost of indexation. Subsequently, the 4IM can be a sufficient basis to represent the different relations that might exist between event features. Therefore, we propose to compute the similarity between every pair of events by calculating the four intersection of the interior and the boundary sets of their similar features.

III.3 The 4-Intersection Model

In this section, we show how we adapt the 4-IM to represent spatial, temporal and social relations. We use a unified representation based on the intersection sets (*IS*) and an intersection matrix (*IM*).

III.3.1 Intersection Set (*IS*)

Following the 4IM, we build our meta-model on the notion of a feature *F*. In this study, *F* could be a spatial, a temporal, or a social feature. For the purpose of homogeneity, we define the following notions and properties of a feature *F* as follows:

- The interior \mathbf{I}^F : is the core of *F*. It is never empty ($\mathbf{I}^F \neq \emptyset$).

- The boundary \mathbf{B}^F : contains other elements related to F without any intersection with I^F ($I^F \cap B^F = \emptyset$). B^F may be empty.
- The tolerance ε : defines the distance or condition that separates I^F from B^F .

The value of ε depends on the (meta)data distribution in the social space of the user at hand. It must take into account the three granularities pre-determined in the pre-processing step of our Event Detection algorithm. Selecting a suitable threshold (for time, location and persons) is important for obtaining accurate (topological) relations. A method of selecting the value of the threshold will be discussed in a future work. In this work, we define it manually.

I^F and B^F constitute what we call the intersection sets of F . In what follows, we define the intersection matrix of two intersection sets between two events. Then, we describe how to define the interior/border for each feature of events: spatial, temporal and social.

III.3.2 Intersection Matrix (IM)

In order to compute relations or links between events e_1 and e_2 , an intersection matrix IM is formed for each corresponding pair-wise event features F_1 and F_2 . It contains the pair-wise intersection results between all possible combinations of interiors and boundaries. IM can be represented as follows:

$$IM_{(F_1, F_2)} = \mathcal{N}_{(F_1, F_2)} = \begin{pmatrix} I_{F_1} \cap I_{F_2} & I_{F_1} \cap B_{F_2} \\ B_{F_1} \cap I_{F_2} & B_{F_1} \cap B_{F_2} \end{pmatrix}$$

where:

- I^{F_1} and I^{F_2} : represent the interiors of the given feature related to e_1 and e_2 respectively;
- B^{F_1} and B^{F_2} : represent the boundaries of the given feature related to e_1 and e_2 respectively.

Each matrix cell has a binary value of 0 whenever the intersection between sets is empty, and 1 otherwise. Each relation between two events' features is thus represented by a binary value. Each computed relation between two features is mutually exclusive: there is *one and only one*

relation between any two features [189]. For example, two events cannot be *spatially* disjoint and equal simultaneously.

IV. Basic Relations between Events

Following our event social space, two events e_1 and e_2 are at least related by three basic relations: spatial, temporal and social. We explain each of these relations in this section. First, we define the interior and the boundary characteristics derived from the spatial, temporal and social information of an elementary event. After, we define the corresponding intersection matrix.

In order to illustrate relations described below, let us take the example of `Lisa`. We select two events shared on her social network.

1. Trip to *Stockholm*:

She made to a trip with her colleagues to *Sweden* in *May 2014*, as described in the trip scenario in Chapter 4.

2. Trip to *Lebanon*:

At the end of *May* of the same year, `Lisa` went back to her home country *Lebanon* to visit her family with her two brothers (`Bouli` and `Elie`).

`Lisa` has two social circles: Family and Colleagues as depicted in Figure 51.

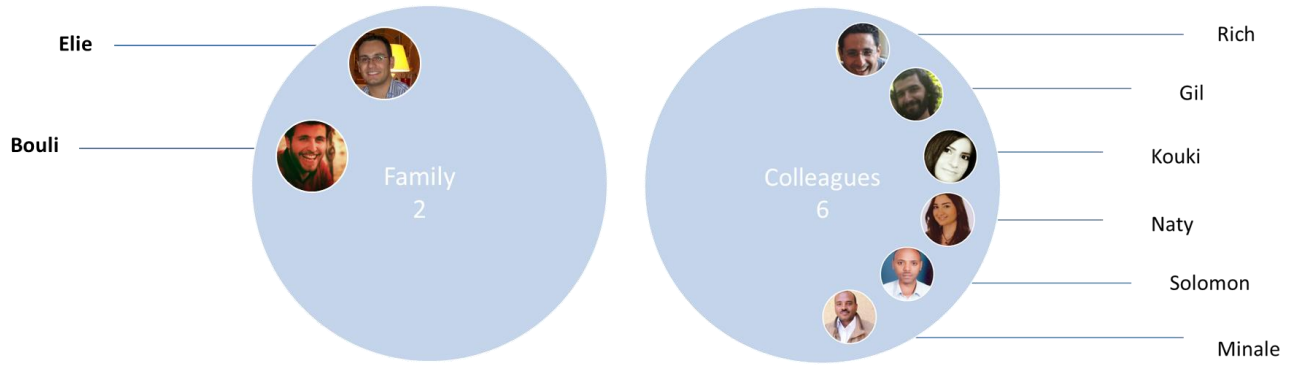


Figure 51: Lisa's social connections. Two social circles: Family: Elie, Bouli; Colleagues: Rich, Gil, Kouki, Naty, Solomon, and Minale.

We assume that Lisa has strong ties with her two family members (Elie, Bouli) and with four of her colleagues: Rich, Gil, Kouki and Naty.

IV.1 Spatial Relations

IV.1.1 Spatial Intersection Sets

Let L_1 and L_2 denote the respective spatial features of e_1 and e_2 . The definitions of their spatial intersection sets are given below:

- $\Delta'_{L1} \subset \Delta_L$ and $\Delta'_{L2} \subset \Delta_L$: represent the bounding polygons in which e_1 and e_2 took place respectively. It includes the set of geographical location names (i.e., $e_1.meta.L$ and $e_2.meta.L$);
- $\Delta''_{L1} \subset \Delta_L$ and $\Delta''_{L2} \subset \Delta_L$: represent the spatial neighborhood of e_1 and e_2 respectively defined as: $f_N(L, \varepsilon) - L$ where f_N is a function adding a certain distance ε to each side of

the bounding polygon(s) of each event. It includes the names of these geographic neighbors.

IV.1.2 Spatial Intersection Matrix

To identify the spatial relation between e_1 and e_2 , we define their intersection matrix with respect to L as follows:

$$IM_{(L_1, L_2)} = \begin{pmatrix} \Delta'_{L_1} \cap \Delta'_{L_2} & \Delta'_{L_1} \cap \Delta''_{L_2} \\ \Delta''_{L_1} \cap \Delta'_{L_2} & \Delta''_{L_1} \cap \Delta''_{L_2} \end{pmatrix}$$

Each element of the 2x2 matrix corresponds to the intersection product between Δ'_{L_1} , Δ'_{L_2} , Δ''_{L_1} or Δ''_{L_2} , i.e., either 0 or 1 depending on whether it is empty or not.

IV.1.3 Topological Spatial Relations

Figure 52 shows the 4-intersection-matrix and the graphical description of the 6 possible topological relations between two spatial intersection sets. The definitions of these relations are given below.

- *disjoint*: The spatial boundaries and interiors do not intersect.
- *equal*: The two events have the same spatial boundary and interior.
- *meet*: The spatial boundaries intersect but the interiors do not intersect.
- *contains*: The spatial interior and boundary of one event is completely contained in the spatial interior of the other event.
- *inside*: The opposite of *contains*.
- *intersects*: The spatial boundaries and interiors of the two events intersect.

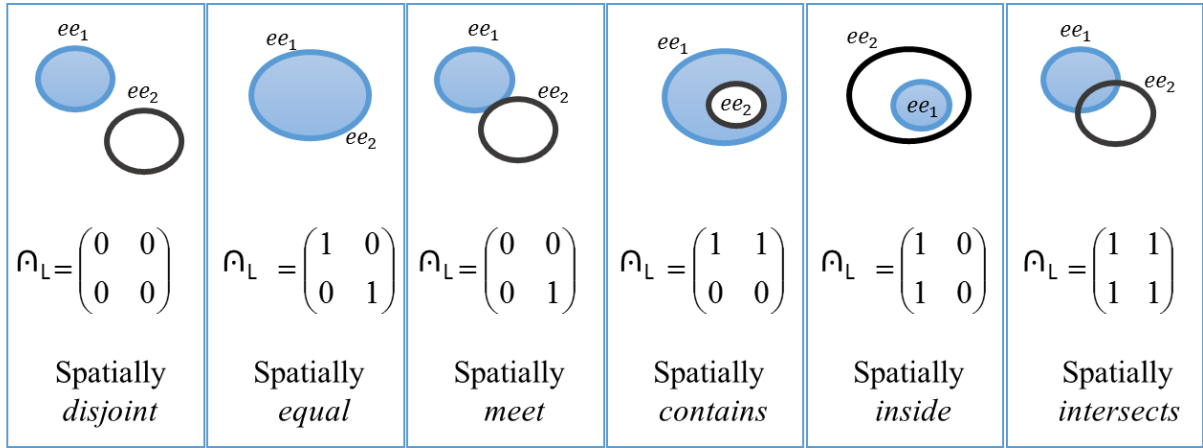


Figure 52: Set of the six possible spatial relations between events (with their graphical description and their corresponding intersection matrices).

IV.1.4 Illustration

We illustrate this on the example given at the beginning of this section: e_1 is the trip to Sweden, and e_2 is the trip to Lebanon. We fix the location threshold to *country*. Δ' and Δ'' in Table 10 constitute the spatial intersection sets.

Table 10: Spatial intersection sets of two events of the example.

\mathbf{IS}_{spa}	e_1	e_2
Δ'	Sweden	Lebanon
Δ''	Norway, Finland, Denmark	Syria, Israel

The intersections between these intersection sets are empty. Then, the intersection matrix will be written as follows:

$$\mathbf{IM}_{(L_1, L_2)} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

The result of \cap_L is read out row by row. For example, the above matrix results in 0000. Therefore, we can say that these two events are *spatially disjoint*.

IV.2 Temporal Relations

IV.2.1 Temporal Intersection Sets

Let L_1 and L_2 denote the respective temporal features of e_1 and e_2 . The definitions of their temporal intersection sets are given below:

- $\Delta'_{T1} \subset \Delta_T$ and $\Delta'_{T2} \subset \Delta_T$: represent the date/time interval in which e_1 and e_2 took place respectively. For instance, Δ'_T is situated between $[t_{tr}, t_{br}] \subseteq \Delta_T$, where:
 - tr : denotes the trigger data/time of the event extracted from the first photo published in e ($\forall i \in \Delta', t_i \geq e.meta.tr$), and
 - br : denotes the break point of an event derived from the last photo published in e ($\forall i \in \Delta', t_i \leq e.meta.br$).

Δ'_{T2} is obtained by the same methodology using event trigger and break.

- $\Delta''_{T1} \subset \Delta_T$ and $\Delta''_{T2} \subset \Delta_T$: represent the temporal neighborhood of e_1 and e_2 respectively obtained by adding a time threshold ε to the trigger and the break points of each event.

Formally, the temporal boundary is represented as:

$$\Delta''_T = [e.meta.tr - \varepsilon, e.meta.tr] \cup [e.meta.br, e.meta.br + \varepsilon]$$

Δ''_{T2} is also calculated in the same manner.

IV.2.2 Temporal Intersection Matrix

To identify the temporal relation that exists between e_1 and e_2 , we define the intersection matrix as follows:

$$IM_{(T_1, T_2)} = \begin{pmatrix} \Delta'_{T1} \cap \Delta'_{T2} & \Delta'_{T1} \cap \Delta''_{T2} \\ \Delta''_{T1} \cap \Delta'_{T2} & \Delta''_{T1} \cap \Delta''_{T2} \end{pmatrix}$$

In order to calculate the elements of this matrix, we follow the same procedure described previously for the spatial feature. The intersection product is computed between Δ'_{T1} , Δ'_{T2} , Δ''_{T1} or Δ''_{T2} .

IV.2.3 Topological Temporal Relations

Figure 53 shows the possible 4-intersection-matrix and the graphical description for the 6 relations between two intervals. The definitions of these relations are given below.

- *Disjoint*: The temporal boundaries and interiors do not intersect.
- *Equal*: The two events have the same temporal boundary and interior.
- *Meet*: The temporal boundaries intersect but the interiors do not intersect.
- *Contains*: The temporal interior and boundary of one event is completely contained in the temporal interior of the other event.
- *Inside*: The opposite of *contains*.
- *Intersects*: The temporal boundary of each event intersects with the temporal interior of the other event.


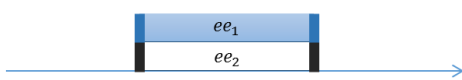

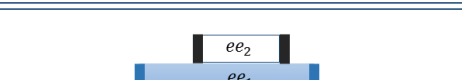
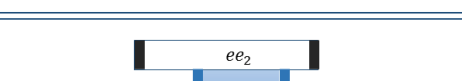
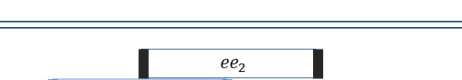
	$\cap_{\mathcal{T}} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$	Temporally <i>disjoint</i>
	$\cap_{\mathcal{T}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	Temporally <i>equal</i>
	$\cap_{\mathcal{T}} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$	Temporally <i>meet</i>
	$\cap_{\mathcal{T}} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$	Temporally <i>contains</i>
	$\cap_{\mathcal{T}} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$	Temporally <i>inside</i>
	$\cap_{\mathcal{T}} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$	Temporally <i>intersects</i>

Figure 53: Set of the six possible temporal relations between events (with their graphical description and their corresponding 4-intersection matrices).

IV.2.4 Illustration

Again, we illustrate this on the example given at the beginning of this section: e_1 is the trip to *Sweden*, and e_2 is the trip to *Lebanon*. Δ' and Δ'' in Table 11 constitute the temporal intersection sets.

Table 11: Temporal intersection sets of two events of the example.

$\mathbf{IS_{tmp}}$	e_1	e_2
Δ'	May	May
Δ''	April, June	April, June

The interiors of e_1 and e_2 intersect. The same can be seen for the boundaries. Then, the intersection matrix will be written as follows:

$$\mathbf{IM}_{(L_1, L_2)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

The result of \cap_T is read out row by row. For example, the above matrix results in 1001. Therefore, we can say that these two events are *temporally equal*.

IV.2.5 Spatial vs. Temporal Relations

Figure 52 and Figure 53 show that the 4-intersections are the same for spatial and temporal relations, except for the *intersect* relation. This is discussed in the literature [188] for line-line relations and region-region relations. Line-line relations are identical to our temporal relations, whereas region-region relations are identical to our spatial relations. According to [188], the difference is due to the fact that regions have connected boundaries, while lines have disconnected boundaries. This fact can be applied to the spatial and temporal relations as well.

IV.3 Social Relations

IV.3.1 Social Intersection Sets

Let L_1 and L_2 denote the respective social features of e_1 and e_2 . The definitions of their social intersection sets are given below:

- $\Delta'_{s1} \subseteq \Delta_S$ and $\Delta'_{s2} \subseteq \Delta_S$: represent the set of people participating in e_1 and e_2 respectively (i.e., $e_1.meta.S$ and $e_2.meta.S$)
- $\Delta''_{s1} \subseteq \Delta_S$ and $\Delta''_{s2} \subseteq \Delta_S$: represent the social neighborhood of e_1 and e_2 respectively. Δ''_{s1} and Δ''_{s2} contains the set of people that are not participating to e_1 and e_2 respectively but belonging to the same groups of those who are participating and having a social tie strength with u_0 larger than a predefined threshold $\varepsilon \in [0, 1]$. Formally, the social boundary is represented as:

$$\Delta''_S = \bigcup_{i=0}^{|\Delta_S|-1} u_i \in \Delta_S - e.meta.S / \left(f_1(u_i) \in \bigcup_{\forall v \in e.meta.S} f_l(v) \right) \wedge (D_{tie}(u_0, u_i) \leq \varepsilon)$$

where:

- $f_1 = SSN_{u_0}.f_1$ is the association function used to identify the relationship type between u_0 and a given user in SSN_{u_0} .
- D_{tie} is a distance function that returns the social tie between two users. It usually comes in three varieties: strong, weak or absent. In our work, we describe users with only either strong or weak tie relationships with u_0 . More details about social ties can be found in [190-193].

Δ''_{s2} is calculated in the same manner.

We illustrate the social intersection set by referring to the example of `Lisa`. We can see the following in Figure 54.

- In the inner circle: the set of people that participated with u_0 in the event. This is the social event interior;

- In the outer circle: the names of the colleagues of u_0 (since Kouki and Naty are linked by the *colleague* relationship with u_0), who have a strong social tie with u_0 . This is the social event boundary;
- In the expanding circle: the rest of the colleagues. In this study, we do not consider the exterior to determine the intersection between features of events.

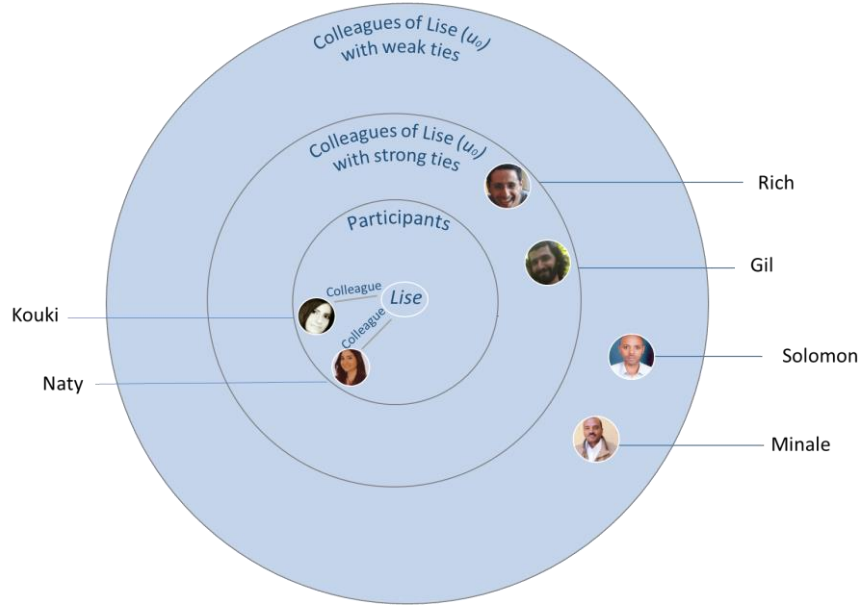


Figure 54: Illustration of event social interior and boundary.

IV.3.2 Social Intersection Matrix

Let S_1 and S_2 denote the respective social features of e_1 and e_2 . To identify the social relation that exists between them, we define the intersection matrix as follows:

$$IM(S_1, S_2) = \begin{pmatrix} \Delta'_{s1} \cap \Delta'_{s2} & \Delta'_{s1} \cap \Delta''_{s2} \\ \Delta''_{s1} \cap \Delta'_{s2} & \Delta''_{s1} \cap \Delta''_{s2} \end{pmatrix}$$

The 4-intersections are then computed between Δ'_{s1} , Δ'_{s2} , Δ''_{s1} and Δ''_{s2} .

IV.3.3 Topological Social Relations with Examples

Now, we give short examples illustrating each social relation. We consider a set of events shared on the social network of `Lisa`. She and her colleagues participate in the same events together all the time and share photos of those moments online. However, each event includes different people.

We show the possible social relations that can link any two events (based on topological relations). Then, we compute the intersection matrix that represents each relation. These relations are valid when the event contains people from many groups (e.g., a birthday party with friends and family). For simplicity, we consider here events with individuals from only a single social group.

IV.3.3.1 Socially *Disjoint* relation

We say that two events e_1 and e_2 are “*socially disjoint*” if the people who participated in e_1 are totally different from those who participated in e_2 , and do not belong to the same group in the social network of u_0 .

To illustrate this relation, we consider two events e_1 and e_2 which are detected in the social network of `Lisa`: in one event, `Lisa` was with three of her colleagues (`Rich`, `Gil` and `Naty`), and in the other, she was with one of her family members (`Bouli`).

We recall that Δ' and Δ'' denote the social event interior and exterior respectively. The “*socially disjoint*” relation has the following binary representation: 0000.

IS_{soc}	e_1	e_2		Social relation
Δ'	Rich, Gil, Naty	Bouli	0000	<i>Disjoint</i>
Δ''	Kouki	Elie		

IV.3.3.2 Socially *Meet* relation

We say that two events e_1 and e_2 “*socially meet*” if the set of participants of one event is different from the others, but both sets belong to only one and same group in the social network of u_0 . In other terms, two events of u_0 , belonging to two groups (one with colleagues and the other with family members), cannot socially meet.

Our meta-model allows us to detect five possible cases with the socially meet relation:

Case 1: Participants of both events have *strong* ties with u_0 .

Case 2: Participants of both events have *weak* ties with u_0 .

Case 3: Participants of one event have *weak* ties with u_0 , and the other event includes *some* people that have *strong* ties with u_0 .

Case 4: Participants of one event have *weak* ties with u_0 , and the other event includes *all* people that have *strong* ties with u_0 .

Case 5: Both events include people that have *strong* ties with u_0 and also people that have *weak* ties with u_0 .

We illustrate these five cases with tables in which we compare the social interior/boundary corresponding to e_1 and e_2 , and compute the corresponding intersection matrix.

Case 1: Consider two events e_1 and e_2 which are detected in the social network of `Lisa`. Each event includes *distinct* groups of colleagues who share *strong* tie strength with her.

IS_{soc}	e_1	e_2		Social relation
Δ'	Rich, Gil	Kouki, Naty	0110	strong- <i>Meet</i> -strong
Δ''	Kouki, Naty	Rich, Gil		

Case 2: Consider two events e_1 and e_2 which are detected in the social network of `Lisa`. Each event includes *distinct* groups of colleagues who share *weak* tie strength with her.

IS _{soc}	e_1	e_2		Social relation
Δ'	Minale	Solomon	0001	weak- <i>Meet</i> -weak
Δ''	Rich, Gil, Kouki, Naty	Rich, Gil, Kouki, Naty		

Case 3: Consider two events e_1 and e_2 which are detected in the social network of Lisa. Each event includes *distinct* groups of colleagues. One event includes some of the colleagues who share *strong* tie strength with u_0 . The other includes colleagues who share *weak* tie strength with u_0 .

IS _{soc}	e_1	e_2		Social relation
Δ'	Rich, Gil	Solomon	0101	strong- <i>Meet</i> -weak
Δ''	Kouki, Naty	Rich, Gil, Kouki, Naty		

Or

IS _{soc}	e_1	e_2		Social relation
Δ'	Solomon	Rich, Gil	0011	strong- <i>Meet</i> -weak
Δ''	Rich, Gil, Kouki, Naty	Kouki, Naty		

Case 4: Consider two events e_1 and e_2 which are detected in the social network of Lisa. Each event includes *distinct* groups of her colleagues. One event includes *all* the colleagues who share *strong* tie strength with Lisa. The other includes colleagues who share *weak* tie strength with u_0 . The same intersection matrix is obtained whether the second group has *all* or some of the colleagues who share *weak* tie strength with u_0 . We explain this by the fact that people with *weak* tie strength will never appear in the boundary set. This can be seen in the following two examples, which have the same intersection matrix.

IS _{soc}	e_1	e_2		Social relation
Δ'	Minale, (Solomon)	Rich, Gil, Kouki, Naty	0010	weak- <i>Meet</i> -ALL strong
Δ''	Rich, Gil, Kouki, Naty			

Or

IS _{soc}	e_1	e_2		Social relation
Δ'	Rich, Gil, Kouki, Naty	Minale, (Solomon)	0100	ALL strong- <i>Meet</i> -weak
Δ''		Rich, Gil, Kouki, Naty		

Case 5: Consider two events e_1 and e_2 which are detected in the social network of Lisa. Each event includes *distinct* groups of colleagues who share *strong* tie strength and also *weak* tie strength with her.

IS _{soc}	e_1	e_2		Social relation
Δ'	Rich, Solomon	Gil, Minale	0111	<i>Meet</i>
Δ''	Gil, Kouki, Nat	Rich, Kouki, Naty		

As a result, we can see from these examples that the “*socially disjoint*” relation has seven binary representations: 0110, 0001, 0101, 0011, 0010, 0100, and 0111.

IV.3.3.3 Socially *Equal* relation

We say that two events e_1 and e_2 are “*socially equal*” if the people who participated in the two events are exactly the same. Our meta-model allows to associate the “*socially equal*” relation with two representations, depending whether the boundary sets are empty or not. Boundary sets are empty when events include all people of one group who share a strong tie strength

with u_0 . Otherwise, boundary sets are not empty. To illustrate this, we consider the following two cases:

Case 1: Lisa had two events e_1 and e_2 both with two of her colleagues Rich and Gil.

IS_{soc}	e_1	e_2		Social relation
Δ'	Rich, Gil	Rich, Gil	1001	Equal
Δ''	Kouki, Naty	Kouki, Naty		

Case 2: Consider two events e_1 and e_2 which are detected in the social network of Lisa, where both events include all the colleagues of Lisa, as shown below.

IS_{soc}	e_1	e_2		Social relation
Δ'	Rich, Gil, Kouki, Naty	Rich, Gil, Kouki, Naty	1000	Equal ALL
Δ''				

IV.3.3.4 Socially *Inside* relation

We say that an event e_1 is “*socially inside*” another event e_2 if the set of people involved in e_1 is completely contained in the set of people involved in e_2 . Here also, our meta-model can associate the “*socially inside*” relation with two representations, depending whether the boundary sets are empty or not. To illustrate this, we consider the two cases:

Case 1: Consider two events e_1 and e_2 which are detected in the social network of Lisa, where e_1 includes a subset of the people involved in e_2 .

IS_{soc}	e_1	e_2		Social relation
Δ'	Kouki, Naty	Gil, Kouki, Naty	1011	<i>Inside</i>
Δ''	Rich, Gil	Rich		

Case 2: Consider two events e_1 and e_2 which are detected in the social network of Lisa, where e_1 includes a subset of the people involved in e_2 . The only difference with *Case 1* is that e_2 includes *all* the colleagues of Lisa who share a strong tie strength with her.

IS_{soc}	e_1	e_2		Social relation
Δ'	Kouki, Naty	Rich, Gil, Kouki, Naty	1010	<i>Inside ALL</i>
Δ''	Rich, Gil			

IV.3.3.5 Socially *Contains* relation

We say that an event e_1 “*socially contains*” another event e_2 if the set of people involved in e_1 comprises all the elements in the set of people involved in e_2 . The “*socially inside*” relation has two binary representations: 1100 or 1101, depending whether the boundary sets are empty or not. When boundary sets are empty, the “*socially contains*” relation will be 1100. Otherwise, it will be 1101. It is the opposite of “*socially contains*”. Therefore, we consider the same examples described for the “*socially contains*” relation, but reversed.

Case 1: Consider two events e_1 and e_2 which are detected in the social network of Lisa, where e_1 includes a subset of the people involved in e_2 .

IS_{soc}	e_1	e_2		Social relation
Δ'	Gil, Kouki, Naty	Kouki, Naty	1101	<i>Contains</i>
Δ''	Rich	Rich, Gil		

Case 2: Consider two events e_1 and e_2 which are detected in the social network of Lisa, where e_1 includes a subset of the people involved in e_2 . The only difference with *Case 1* is that e_2 includes *all* the colleagues of Lisa who share a strong tie strength with her.

IS _{soc}	e_1	e_2		Social relation
Δ'	Rich, Gil, Kouki, Naty	Kouki, Naty	1100	<i>Contains ALL</i>
Δ''		Rich, Gil		

IV.3.3.6 Socially *Intersect* relation

We say that an event e_1 “*socially intersects*” with another event e_2 if there is an intersection in the set of people involved in e_1 and e_2 . The “*socially intersect*” relation has two representations using our meta-model, depending on the cardinality of *IS*.

As we can see in *Case 2*, each event includes three colleagues having strong tie strength with u_0 . Two of them are common (Rich and Gil). Since there are a total of four colleagues who have strong tie with u_0 , the boundary will contain only one (and different) colleague in each event. Therefore, there is no intersection between boundaries.

This is the only difference with *Case 1* where more than one colleague appear in the boundaries. Consequently, the intersection between boundaries is not empty and yields a binary representation of the form: 1111, instead of 1110 in *Case 2*.

Case 1: Lisa had two events: e_1 with two of her colleagues Kouki and Naty; and e_2 with two of her colleagues Gil and Naty.

IS _{soc}	e_1	e_2		Social relation
Δ'	Kouki, Naty	Gil, Naty	1111	<i>Intersect</i>
Δ''	Rich, Gil	Rich, Kouki		

Case 2: Lisa had two events with *three* of her colleagues: e_1 with Rich, Gil and Naty; and e_2 with Rich, Gil and Kouki.

IS_{soc}	e_1	e_2		Social relation
Δ'	Rich, Gil, Naty	Rich, Gil, Kouki	1110	<i>Intersect ALL</i>
Δ''	Kouki	Naty		

Figure 55 shows the 4-intersection-matrix and the graphical description of the 6 possible topological relations between two social intersection sets.




e_1		e_2		$\cap_s = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$	Socially <i>disjoint</i>
e_1		e_2		$\cap_s = \begin{pmatrix} 1 & 0 \\ 0 & * \end{pmatrix}$	Socially <i>equal</i>
e_1		e_2		$\cap_s = \begin{pmatrix} 0 & 0 \\ 1 & * \end{pmatrix} \text{ ou } \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \text{ ou } \begin{pmatrix} 0 & 1 \\ * & * \end{pmatrix}$	Socially <i>meet</i>
e_1		e_2		$\cap_s = \begin{pmatrix} 1 & 0 \\ 1 & * \end{pmatrix}$	Socially <i>inside</i>
e_1		e_2		$\cap_s = \begin{pmatrix} 1 & 1 \\ 1 & * \end{pmatrix}$	Socially <i>intersects</i>

Figure 55: Set of the six possible social relations between events (with their graphical description and their corresponding intersection matrices).

IV.3.4 Spatial vs. Temporal vs. Social Relations

We summarize this in Table 12 where we show the relations' representations.

Table 12: The binary and hexadecimal values of the three basic relations (spatial, temporal and social)

Hexadecimal values	Binary values	Spatial relation	Temporal relation	Social relation
0	0000	disjoint	disjoint	disjoint
1	0001	meet	meet	weak- <i>Meet</i> -weak
2	0010			weak- <i>Meet</i> - <i>ALL</i> strong
3	0011			weak- <i>Meet</i> -strong
4	0100			<i>ALL</i> strong- <i>Meet</i> -weak
5	0101			strong- <i>Meet</i> -weak
6	0110			strong- <i>Meet</i> -strong
7	0111			meet
8	1000			equal <i>ALL</i>
9	1001	equal	equal	equal
A	1010	inside	inside	inside <i>ALL</i>
B	1011			inside
C	1100	contains	contains	contains <i>ALL</i>
D	1101			contains
E	1110		intersects	intersects <i>ALL</i>
F	1111	intersects		intersects

We can see from Table 12, and as we mentioned before, each spatial or temporal relation has one binary value, and thereby one hexadecimal value. However, for social relations, we can define ranges of values as follows:

disjoint: 0; *meet*: 1-7; *equal*: 8-9; *inside*: A-B; *contains*: C-D; *intersects*: E-F.

V. Event Constraints

Our approach goes beyond identifying spatial, temporal and social relations between events. Each of the spatial (R^L), the temporal (R^T) and the social (R^S) relation can result in one of the 6 different types of topological relations: *disjoint*, *equal*, *meet*, *contains*, *inside*, or *intersects*. In this section, we propose to group all possible combinations of these three relations. This results in a total of 6^3 combinations. However, not all combinations are valid, as they can lead to impossible cases because of spatial, temporal and social constraints that we discuss here.

V.1 Time-based Constraints

Given two elementary events e_1 and e_2 that are *temporally equal*, participants cannot be the same if e_1 and e_2 take place in two different locations. This *Temporality Constraint* is based on our knowledge that a person cannot physically be in two places at the same time. Thus, we assume that a person cannot be involved in more than one event at the same time, similarly to [91, 125, 194]. For example, `Lisa` is in *Sweden* or in *Lebanon*, therefore it is not possible that `Lisa` participates in one event in *Sweden* and another in *Lebanon* simultaneously.

V.2 Location-based Constraints

Let e_1 and e_2 denote two events that involve one or more persons in common, occur at different locations, and a time bound $T \in \mathbb{N}$ representing the interval between e_1 and e_2 . The *Reachability Constraint* ensures that e_1 and e_2 cannot be located at a distance greater than D_{max} , where D_{max} represents the maximum distance that can be travelled in T by any mode of transport (e.g., roads, railways, inland waterways or airports, etc.).

V.3 Social-based Constraints

In social networks, people are connected by specific relationship (or link) categories like friends, colleagues or family. We assume that a user can have *only one* relationship with another user within a single social network. For instance, `Lisa` and `Bouli` are either relatives or colleagues. This assumption would distinguish event categories using the relationship types that exist in the social network (e.g., family events, professional events, etc.).

V.4 Pruning Relations Combinations

As we mentioned before, when examining event relations between i) people, ii) places and iii) time separately, it is necessary to check if the intersection of these three relations is also valid. The time, location and social-based constraints are used for pruning the set of possible combinations of relations, so that only the valid ones are selected. All the possible cases are listed in Table 13, where:

- The rows list the 6 temporal relations (R^T), and
- The columns list the 6 spatial relations (R^L).

The value of each cell corresponds to the names of the social relation when the combination of the three relations is possible to have in real-life events. Let us consider the first term in Table 13. It contains the set of social relations that are possible when two events are *disjoint* spatially and temporally. In this case, the 6 social relations are all possible (D, M, E, C, In, It).

Table 13: The possible combinations of the three relations:

disjoint (D), *meet* (M), *equal* (E), *contains* (C), *inside* (In) and *intersects* (It)

$R^T \backslash R^L$	D	M	E	C	Is	It
D	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It
M	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It
E	D, M	D, M	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It
C	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It
Is	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It
It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It	D, M, E, C, In, It

Table 13 shows that we can reduce the $6^3 = 216$ possible cases to 3 groups:

Group 1: There is no constraint on the combinations of relations when the temporal relation between events is *disjoint* or *meet* (as shown in the rows 2 and 3 of the above table). When events do not occur at the same time, they can involve whoever and wherever. Therefore, the spatial and the social relations can be any one of the aforementioned relations (D, M, E, C, In or It).

Group 2: Based on the time constraint mentioned above and using the social constraint imposing the unicity of relationships between users, we can show that there are some impossible cases when two events are *temporally equal* (as shown in the fourth row of the

above table). In this case, when the spatial relation between e_1 and e_2 is *disjoint* or *meet*, the social relation can be only *disjoint* or *meet*. We formulate this constraint as follows:

$$(D \mid M) \quad (E) \quad (D \mid M)$$

where:

- The terms in the first parentheses correspond to the spatial relations (*disjoint*, *meet*) separated by the pipe (|) sign to indicate “or”;
- The terms in the second parentheses correspond to the temporal relation *equal*;
- The terms in the third parentheses correspond to the possible social relations (*disjoint*, *meet*) separated by the pipe (|) sign indicating “or”.

Group 3: The reachability constraint is necessary for some cases to ensure that there is enough time to get from one event's location to another event's location, when the two events take place in different locations (as shown in the rows 4, 5 and 6 of the above table). Therefore, when the social relation is neither *disjoint* nor *meet* (because of the social constraint), the reachability constraint has to be checked. If the time difference and the distance between the two events satisfy the reachability constraint, the temporal relations can be any one of the aforementioned relations (D, M, E, C, In or It). We formulate this constraint as follows:

$$(\overline{E}) \quad (D \mid M \mid E \mid C \mid In \mid It).RC \quad (\overline{D} \& \overline{M})$$

where:

- The terms in the first parentheses correspond to the spatial relations *not equal*;
- The terms in the second parentheses correspond to the possible temporal relations separated by the pipe (|) sign to indicate “or”. It is indicated that the reachability (RC) constraint has to be checked;
- The terms in the third parentheses correspond to the social relations (*not disjoint*, *not meet*) separated by “&” indicating “neither”;

VI. Computing Links between Events

As we mentioned before, every two events are related by three relations: one spatial, one temporal and one social relation. Thus, event relations can be represented as:

$$R(e_i, e_j): (R^L, R^T, R^S)$$

where:

- R^L : is an exclusive spatial relation that e_i has with e_j , and corresponding to the l value of the Δ_L^R axis of the social-R Space.
- R^T : is an exclusive temporal relation that e_i has with e_j , and corresponding to the t value of the Δ_T^R axis of the social-R space.
- R^S : is an exclusive social relation that e_i has with e_j , and corresponding to the s value of the Δ_S^R axis of the social-R space.

Now that we have the three exclusive relations between two events, we can define their Hamard Intersect product.

VI.1 From Intersection Set to Hamard Intersect Product

Let IS_{e1} and IS_{e2} denote the intersection sets of two elementary events e_1 and e_2 . The Hamard product [195], denoted by \circ , is the element-by-element product of the two matrices. By definition, the result is a matrix R as shown below:

$$R(e_i, e_j) = \circ \cap (IS_{e1}, IS_{e2}) = \begin{pmatrix} IS_{e1} \cdot \Delta_L \cap IS_{e2} \cdot \Delta_L \\ IS_{e1} \cdot \Delta_T \cap IS_{e2} \cdot \Delta_T \\ IS_{e1} \cdot \Delta_S \cap IS_{e2} \cdot \Delta_S \end{pmatrix}$$

where:

- $IS_{e1} \cdot \Delta_L \cap IS_{e2} \cdot \Delta_L$: denotes the spatial relation between e_1 and e_2 . It is obtained by applying the intersection matrix between IS_{e1} and IS_{e2} with respect to the location feature L.

- $IS_{e1}.\Delta_T \cap IS_{e2}.\Delta_T$: denotes the temporal relation between e_1 and e_2 . It is obtained by applying the intersection matrix between IS_{e1} and IS_{e2} with respect to the temporal feature T.
- $IS_{e1}.\Delta_S \cap IS_{e2}.\Delta_S$: denotes the social relation between e_1 and e_2 . It is obtained by applying the intersection matrix between IS_{e1} and IS_{e2} with respect to the social feature S.

VI.2 From Hamard Intersect Product to Binary Triplet

Suppose two events e_1 and e_2 occur in two different sites at the same time and include two different groups of colleagues in SSN_{u0} . Then, the three intersection matrices are as follows:

$$\cap_L = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} ; \cap_T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} ; \cap_S = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

As we mentioned before, the result of each intersection matrix is read out in our study row by row. This comes down to 12-bit (3×4-bit values) as we have three intersection matrices. Their Hamard product will be:

$$R(e_1, e_2) = \circ \cap (IS_{e1}, IS_{e2}) = \begin{pmatrix} 0000 \\ 1001 \\ 0110 \end{pmatrix}$$

We convert the result of the Hamard product to a binary triplet. It will have the following value:

$$R(e_1, e_2) = 0000 \ 1001 \ 0110$$

This would help later to give a binary representation of complex relations, as explained in the next section.

VII. Complex Relations of Events

Linking events is not simple to compute when it is about personal events because such events represent human experience. Personal events might be connected in multiple complex ways. Thus, the relations between them would be complex and need a variety of aspects to be met simultaneously. Complex relations cannot be determined using only temporal, spatial or social data. Combination of two or three aspects with other semantic descriptors (e.g., topic or thematic information, etc.) may be needed to find the correct relations.

In the following, we present first our form to express complex relations between events. Then, we use some classical ontological relations in order to identify semantic relations between events. The purpose here is to provide human-understandable semantics for events detected in shared photos. For example, this will allow us to find the relations between events and sub-events or sub-collections, and also to find events that cause other events or inherit from other events.

VII.1 Logical Expression

We build an expression in order to define complex relations among events. We represent the expression denoting the relation in terms of logical operators and predicates. This expression is written as:

$$expr ::= [O] (R \mid Predicate \mid expr)$$

where:

- *expr* is the name of the relation;
- *O* is a logical operator (AND, OR, NOT);
- *R* is a contextual relation with one, two or three features (spatial, temporal, social);
- *Predicate* is a predicate that uses user preferences (e.g., her favorite location), relationships between users (e.g., “colleagues”), or event metadata (e.g., “Dijon”).

The choice of this particular expression allows the combination of one or more contextual relations to derive complex relations using logical operators. In addition, predicates can help in modelling complex events because they may provide valuable information to the event linking that is not revealed in the relational social graph. Hence, contextual parameters given by the user as predicates could be potentially useful to obtain additional relations that are meaningful to the user.

VII.2 Ontological Relations

In the following, we introduce some ontological relations between events and represent them as binary triplets. Examples will be given in the following definitions.

VII.2.1 isSameAs Relation

In our context and based on the multi-* property of events described in Chapter 4, the *isSame-as* relation can have four possible values:

- i. The *isSame-as* relation (or the synonymy relation in WordNet [196]) corresponds to the relation between two events with the same space-time and social features. An event e_1 *isSame-as* another event e_2 if e_1 and e_2 are related by the equality relation on the three event aspects: spatially, temporally and socially. Thus, the *isSame-as* relation can be represented as follows:

$$isSame-as ::= (1001 \ 1001 \ 1001) \text{ OR } (1001 \ 1001 \ 1000)$$

such as:

- The first 4-bit 1001 is for the spatial relation (E);
- The second 4-bit 1010 is for the temporal relation (E);
- The third 4-bit 1001 or 1000 are for the social relation (E). We recall that the relation *socially equal* has two representations.

This first representation can be simplified to:

$$isSame-as ::= (1001 \ 1001 \ 100_)$$

For example, it is common that many people take photos using their smartphones during a wedding event. Photos taken in the event by *different creators* may be identified as different elementary events by our event detection algorithm. In essence, these multiple events happened in a single day, at the same exact location and with the same people. Therefore, they are linked by the relation: *isSame-as*.

- ii. In some cases, a wedding event can span *several days* and has to be considered as one event. Therefore, the *isSame-as* relation holds between two elementary events that *temporally meet* even if they are related by the equality relation on only two aspects: spatially and socially. Thus, the *isSame-as* relation will be represented as follows:

$$isSame-as ::= (1001 \ 0001 \ 1001) \text{ OR } (1001 \ 0001 \ 1000)$$

such as:

- The first 4-bit 1001 is for the spatial relation (*E*);
- The second 4-bit 0001 is for the temporal relation (*M*);
- The third 4-bit 1001 or 1000 are for the social relation (*E*);

The second representation can also be simplified to:

$$isSame-as ::= (1001 \ 0001 \ 100_)$$

- iii. The following example illustrates a third particularity of some (wedding) events: they can be scattered at *different sites*, and sometimes people have to move into different cities in the same day. Again, the *isSame-as* relation holds between two elementary events that are related by the equality relation on these two aspects: temporally and socially. Thus, the *isSame-as* relation will be represented as follows:

$$isSame-as ::= (______ \ 1001 \ 1001) \text{ OR } (______ \ 1001 \ 1000)$$

such as:

- The first ____ is for any spatial relation (*not equal*);
- The second 4-bit 1001 is for the temporal relation (*E*);
- The third 4-bit 1001 or 1000 are for the social relation (*E*);

The third representation can also be simplified to:

$$isSame-as ::= (____ 1001 100__)$$

The three representations of *isSame-as* relation enable us to address the issue of multi-*source*, multi-*day*, multi-*site* and multi-*participant* events respectively.

VII.2.2 isSubevent-of Relation

The *isSubevent-of* relation (or the mereological relation in WordNet) reflects how events can be composed of two or more events. It indicates the part-whole hierarchical relations in events. An event e_1 *isSubevent-of* another event e_2 when there is:

- A temporal containment (*inside*), and
- A social equality, intersection or containment (*equal*, *intersects*, *contains* or *inside*).

Thus, the *isSubevent-of* relation can be represented in two forms:

i. *isSubevent-of* ::= (____ 1010 11__) OR (____ 1010 10__) such as:

- The first 4-bit is for the spatial relation;
- The second 4-bit 1010 is for the temporal relation (*In*);
- The third 11__: is to simplify four social relations: 1100 (*C*), 1101 (*C*), 1110 (*It*), 1111 (*It*).

ii. *isSubevent-of* ::= (____ 1010 11__) or (____ 1010 10__) such as:

- The first 4-bit is for the spatial relation;
- The second 4-bit 1010 is for the temporal relation (*In*);
- The third 10__: is to simplify four social relations: 1000 (*E*), 1001 (*E*), 1010 (*In*), 1001 (*In*).

For instance, a keynote session is a sub-event of a big conference. The conference took place at the UPPA University in Bayonne from December 17 to December 19, 2014. The keynote session was scheduled for the first day in one conference room of the University. Over sixty persons attended both the conference and the keynote session. Thus, the following conclusions can be drawn:

- i. Firstly, the location of the session is spatially *inside* the location of the conference.
- ii. Secondly, the date of the session is *within/ inside* the start and end dates of the conference.
- iii. Thirdly, the conference's participants include the *same* participants of the keynote.

Concretely, this *isSubevent-of* relation results in the following binary triplet:

(1001 1010 1000)

VII.2.3 isSubcollection-of Relation

The *isSubcollection-of* relation (or aggregation relation) occurs when a larger event can be viewed as a collection of smaller events. Such events cannot be viewed as sub-events of the larger event because they do not involve the same people. However, they could be of common topics of interest, with an aggregation over time and space. A topic is what the event is about. Recently, several studies have been conducted on event topic detection in social networks [197-201]. We note that in this study we do not address event topic in a particular way. This will be addressed in a dedicated study. Thus, in addition to the topic similarity, an event *isSubcollection-of* a bigger event if we can observe the following:

- Spatial containment (*inside*), and
- Temporal containment (*inside*).

Thus, the *isSubcollection-of* relation is represented as follows:

isSubcollection-of ::= (1001 1010 ____)

isSubcollection-of ::= (1010 1010 ____)

For instance, the “official matches” are a collection of “Fifa World Cup” events. In 2014, the tournament began on 12 June and concluded on 13 July (the date of the championship match)⁹⁴. 12 venues in twelve cities were selected for the tournament, all located in Brazil. Let us consider two games of the Fifa World Cup that were played in a single day: 12 July. The first one was located in São Paulo, while the second one was located in Rio de Janeiro. These games are socially unrelated and cannot be connected to the big event for having common participants. However, they are linked by three relations:

- There is a “inside” relation between the two cities (São Paulo, Rio de Janeiro) and the country where the “Fifa World Cup” event was held (Brazil).
- The temporal relation “inside” holds between the day of the games 12 July and the “Fifa World Cup” event interval (12 June – 13 July).
- They share the common topic of the big event: “soccer”. In this work, we do not explore how to identify event topic.

Therefore, we conclude that the games played in the two different cities are sub collection of the larger soccer event in Brazil.

VII.2.4 caused-by Relation

Given two events e_1 and e_2 , it is possible that e_1 is the cause of e_2 , or that e_2 is the cause of e_1 . This is the causal relation between two events. It helps to answer questions related to *why* a particular event took place. Causal relations require the modeling of causes and effects and should support the integration and use of different causal theories [202, 203]. Therefore, despite the interesting modeling and properties of causal links, we do not address the “why” question in this work. It will be important for future work to study causal relations between events.

⁹⁴ http://en.wikipedia.org/wiki/2014_FIFA_World_Cup#Match_summary

VII.2.5 is-a Relation

Like the inheritance relation in the ontology, the *is-a* relation between events represents the generalization or specialization relation between two events, where one event is a specialized version of another. For instance, the *is-a* relation between *session* and a *keynote session* is a subclass-superclass relation. The inheritance rules are the following:

- The instances of *keynote session* inherit features and properties present in the parent *session*.
- An instance of *session* can be member of only one of its subclasses at all instants of time.

The generalization and specialization relation result in event hierarchy where event types and subtypes are required such as a birthday party, a soccer game, etc. Typically, such relations between events need classifiers that are learned using a set of training images [204]. Such classifiers allow the computation of a decision or a probability that a photo is of a certain known event type. Yet, event types are not involved in our model.

VII.3 Application-based Relations

Some events may not have a common cause or aspect; however, they might be related along some particular aspects that make sense only to the user. Hence, these relations must be defined by the user's preferences using predicates. We consider some examples in order to illustrate this possible extension.

withFriendsGroup: is the relation between two events happening with u_0 ' friends.

$$withFriendsGroup ::= withSameGroup \wedge f_i(e.meta.S) = \text{"friends"}$$

where $withSameGroup ::= (\text{____} \text{ ____ } \text{ ____} 1)$ is the relation between two events happening with the same group of participants in SSN_{u_0} . This means that the social boundaries of the two events must be identical. Similar expressions can be applied to the *withFamilyGroup*, *withCollegeGroup* relations, etc

withAmy: is the relation between two events experienced with *Amy*, one of the friends of u_0 and the users on SSN_{u_0} .

$$withAmy ::= f_l(e.meta.S) \text{ CONTAINS "Amy"}$$

atAirports: is the relation between two events where pictures are taken in airports. For example, when taking a trip, sometimes people start taking photos at the airports, or in the take-off and landing.

$$atAirports ::= f_l(e.meta.L) = \text{"airport"}$$

atHome: is the relation between two home events (e.g., a dinner party at home, etc.).

$$atHome ::= f_l(e.meta.L) = \text{"home"}$$

atWeekends: is the relation between two events occurring at weekends, such as going to saturday parties, or going on a weekend trip.

$$atWeekends ::= f_l(e.meta.T) = \text{"saturday"} \text{ OR } \text{"sunday"}$$

atWeekendsWithFamily: is the relation between two events occurring at weekends with u_0 ' family members.

$$atWeekendsWithFamily ::= withSameGroup \wedge (f_l(e.meta.S) = \text{"family"}) \wedge \\ (f_l(e.meta.T) = \text{"saturday"} \text{ OR } \text{"sunday"})$$

VIII. Conclusion

In this chapter, we have presented our matrix-based approach for identifying relationships between events. We described the different relationships that we are interested in. We first detailed the basic ones which are related to the spatial, temporal and social aspects of events. Then, we showed how to combine the set of basic relations to obtain more complex relations.

As discussed above, event semantization is very important. Figure 56 represents the transitions from multimedia data, to event information, and finally to knowledge (linked events) using our approach.

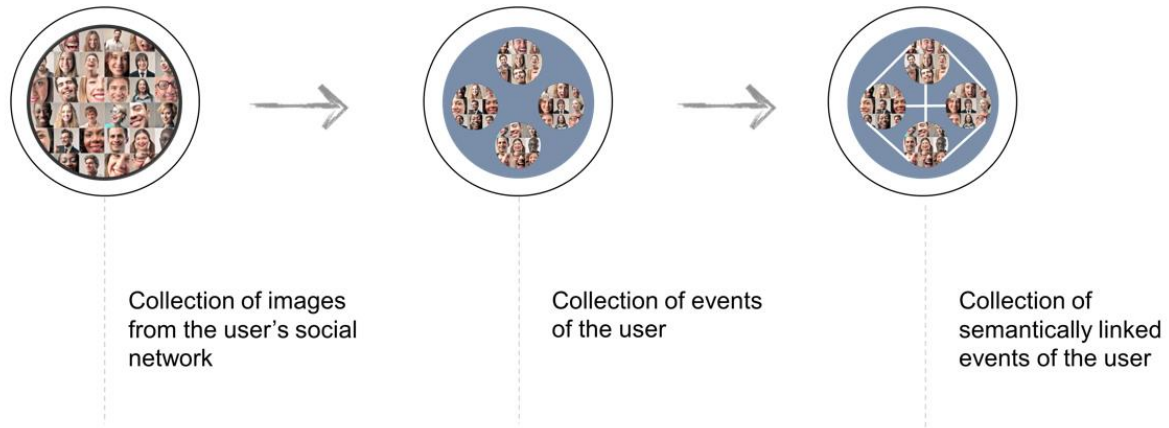


Figure 56: The Event Semantization process.

When detecting events from the collection of images in the user's social network, the issue of information semantics is one of the major challenges. This is what we attempted to address in this chapter, by employing a three-dimensional space drawn from the main characteristics of events (*Who*, *Where*, *When*).

The two main benefits that can be obtained from this approach are:

- The identification of equivalent events (i.e. related by the semantic relation *isSame-as*) to deal with particular events: *multi-source* events, *multi-participant* events, *multi-site* events, and *multi-day* events, as described in the previous chapter.
- The enrichment of events with semantic relations to achieve semantically richer event detection.

In addition, the advantage of our approach is that it allows the homogeneous representation of relationships between various aspects: spatial, temporal and social, and takes into account all

possible combinations of these relations. In addition to the basic relations, we exploited other complex relations. We modeled relations with a binary representation, which is supposed to be easy, flexible and extensible. Therefore, our approach is expected to be able to be applied for defining arbitrary relations that can be defined by the social network user based on his context or his preferences as well.

For optimal expression in future works, each matrix cell can also have one of the following value (but not limited to this):

- The cardinality of the intersection set, instead of only distinguishing empty or non-empty intersections as in the DE-9IM (Dimensional Extended nine-Intersection Model),
- A general concept such as:
 - A common place (CP): represents a location where u_0 spends the most of his time. This information may be provided in the user profile (e.g., home and work locations).
 - An uncommon place (UP): represents a location that is away from CP. It can be determined based on the metric distance of geographical locations ($\Delta g_i = g_i - g_{CP}$).
 - A common face (CF): belongs to a person that often appears in the photos of u_0 or posts photos containing the face of u_0 . The Jaccard coefficient [38] can be applied to capture the degree of co-occurrence of a given person and u_0 .
 - An uncommon face (UF): belongs to a person that has only one occurrence in the same photo with u_0 or has repeated occurrences over a specific period.

Chapter 6: Conclusion

*“The citizens will divide between those
who prefer convenience and those who
prefer privacy.”*

- Niels Ole Finnemann

*A professor and director of NETLAB,
DIGHUMLAB IN DENMARK*

I. Contributions

The study presented in this thesis has mainly been concerned with preserving privacy of anonymous users when publishing multimedia documents online, specifically photos as we explained in **Chapter 1**. We investigated metadata's photos to acquire more information about the users to protect their anonymity. Particularly in this thesis, we are interested in personal information that can be inferred from events shared through photos on online social networks. The contributions of this work are summarized in the following:

Chapter 2 contained the background review, in which we focused on privacy preservation and inference control. We also discussed various definitions of online identity and how it can be constructed using the social networking sites. Then, we investigated the techniques used for online event detection and linking using multimedia content available on social networks.

In **Chapter 3**, we focused on the anonymization of multimedia documents published online that can put at risk the privacy of users and violate their anonymity. In our motivating scenario described in Chapter 1, we showed the need to preserve the identity of anonymous users (like in the case of the previous French Prime Minister François Fillon). To address this, we presented a privacy-preserving constraint, called *de-linkability*, to prevent the leakage of sensitive information (anonymity, pseudonymity, etc.) caused by textual and multimedia content that users produce online. With this constraint, users can be warned of potential violation of their privacy. Then, we provided a sanitizing *MD**-algorithm to enforce *de-linkability* along with a utility function to evaluate the utility of multimedia documents that is preserved after the sanitizing process. Thus, the identity of the user who wishes to remain anonymous is kept private. We developed a prototype to evaluate the identity anonymization problem and the proposed sanitizing algorithm. We experimentally demonstrated how the *de-linkability* could break the link between multimedia documents to be published (i.e., tweets) and any other document accessible to a potential adversary and that can be linked to the user.

In **Chapter 4**, we focused on exploiting photos' metadata and tags shared in online social networks to detect personal events, sometimes of capital importance for privacy protection. To do so, we introduced the Social Space based on the three main aspects of event: temporal, geographical and social, and which are used in the description of our approach. We also defined an *elementary event* and proposed a method to cluster photos shared by social network users or by their friends/contacts in a social network. Our algorithm relies on a pre-processing step that defines the spatio-temporal granularities in order to obtain better quality clusters of events. We developed a prototype called *Foto2Event* to validate and demonstrate the efficiency of our event detection approach presented in this work. We successfully demonstrated that our theoretical analysis is accurate in predicting the appropriate granularities, since it was in good agreement with the experimental observation. We also tested the relevance of our proposal using the MediaEval dataset and compared our results to other algorithms implemented in the Social Event Detection MediaEval Benchmarking Initiative for Multimedia Evaluation⁹⁵.

In **Chapter 5**, we studied relations that exist between events in order to strengthen further the detection of personal information (directly and indirectly). We designed a meta-model to represent various relations in a unified and flexible way. First, we identified *temporal*, *spatial* and *social* relations that exist between any two events based on an extension of the 4-Intersection Model. Then, we computed these basic relations together to produce more complex relations. Our meta-model is also capable to represent other types of relations between events:

- a) Basic relations: temporal, spatial and social relations based on topological relations (*disjoint*, *meet*, *equal*, *contains*, *inside*, and *intersects*),
- b) Ontological relations (e.g., *isSameAs*, *isSubevent*, *isCausedBy*, etc.), and
- c) Application-based relations (e.g., *atHome*, *atAirport*, etc.).

⁹⁵ <http://www.multimediaeval.org/mediaeval2014/sed2014/>

II. Future Works

There are several improvements we intend to make to our existing contributions and validation.

II.1 Improving theoretical Approach

First, we hope to improve our sanitizing algorithm using dedicated techniques that anonymize the visual content of multimedia documents while preserving their semantics and coherence (e.g., Repetitive denoising (RD), Adaptive PRNU denoising (APD), etc.).

Second, more investigations are still needed to improve our event detection approach:

- Capture the metadata of an event (representative moments, locations, and people/groups) by reasoning on the *significant-frequent* and *significant-rare* items (metadata) from one event photos.
- Filling missing metadata of photos from the obtained events and their links, in order to produce better clustering results.
- Explore the “What”, “Why” and “How” aspects of an event, since so far our Social Graph represents the “Who”, “Where” and “When” of an event. This extension is expected to give a more exact analysis (clustering, link computing, and reasoning) from events detected on social networks, and therefore can help to discover the semantics of users’ events.

II.2 Improving Validation

In the future, we plan to enhance our *Foto2Event* prototype by generating tags for photos and relationships between users (e.g., *friends*, *colleagues*, *relatives*, etc.). We intend to provide a public web-based version since it could help to test and validate our obtained results. Finally, we plan to implement and test our model of event relations, and then link it to our sanitizing algorithm to integrate all our proposed algorithms and techniques.

In the following, we present some of the possible research directions.

II.3 New Research Directions

As for new directions in research, we need to quantify the quality of information retrieved from events before using it to enrich users' profiles. Measuring quality requires reliable measures of data accuracy, while taking into account the evolution of (event) information over time (e.g., timeliness, freshness, up-to-dateness, etc.).

Another interesting aspect to consider in the future is the user's engagement in events. In other words, we are interested to measure the user's interest degree in a particular event based on his personal information and his personal events and their links as well (events shared by himself or by his connections):

- For a past event, we will compute the likelihood degree of user's involvement in an event, and infer his presence even if he did not appear in any photo of that cluster.
- For an upcoming event, it would be possible to predict if the user might be interested in an event. If so, we can predict if he will be able to participate or not, by calculating the probability of the user being elsewhere when that event will happen (Event prediction).

Another future direction that we would like to explore is sentiment analysis in events detected in online social networks. We are interested in detecting the feeling of a user in a particular event. This would allow classifying events as "happy", "sad" or "neutral", by mining text accompanying photos, recognizing facial expressions, identifying the event type (e.g., party, birthday, sickness, death, etc.).

References

1. Podesta, J., *Big Data: Seizing Opportunities Preserving Values*, 2014.
2. Malin, B., *Trail re-identification and unlinkability in distributed databases*, 2006, Carnegie Mellon University.
3. Andrews, L., *I know who you are and I saw what you did: Social networks and the death of privacy* 2012: Simon and Schuster.
4. Tavani, H.T. and F.S. Grodzinsky, *Cyberstalking, personal privacy, and moral responsibility*. *Ethics and Information Technology*, 2002. 4(2): p. 123-132.
5. Motahari, S., J.M. Mayer, and Q. Jones, *How did you know that about me? Protecting users against unwanted inferences*. *EAI Endorsed Trans. Security Safety*, 2011. 1: p. e3.
6. Abiteboul, S., B. André, and D. Kaplan, *Managing your digital life*. *Communications of the ACM*, 2015. 58(5): p. 32-35.
7. Agrawal, D. and C.C. Aggarwal, *On the design and quantification of privacy preserving data mining algorithms*, in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* 2001, ACM: Santa Barbara, California, USA. p. 247-255.
8. Clifton, C., et al., *Tools for privacy preserving distributed data mining*. *SIGKDD Explor. Newsl.*, 2002. 4(2): p. 28-34.
9. Evfimievski, A., J. Gehrke, and R. Srikant, *Limiting privacy breaches in privacy preserving data mining*, in *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* 2003, ACM: San Diego, California. p. 211-222.
10. Evfimievski, A., et al., *Privacy preserving mining of association rules*, in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* 2002, ACM: Edmonton, Alberta, Canada. p. 217-228.

11. Pinkas, B., *Cryptographic techniques for privacy-preserving data mining*. SIGKDD Explor. Newsl., 2002. 4(2): p. 12-19.
12. Bertino, E., D. Lin, and W. Jiang, *A Survey of Quantification of Privacy Preserving Data Mining Algorithms*, in *Privacy-Preserving Data Mining*, C. Aggarwal and P. Yu, Editors. 2008, Springer US. p. 183-205.
13. Verykios, V.S., et al., *State-of-the-art in privacy preserving data mining*. SIGMOD Rec., 2004. 33(1): p. 50-57.
14. Heurix, J. and T. Neubauer, *Privacy-preserving storage and access of medical data through pseudonymization and encryption*, in *Proceedings of the 8th international conference on Trust, privacy and security in digital business*2011, Springer-Verlag: Toulouse, France. p. 186-197.
15. Farkas, C. and S. Jajodia, *The inference problem: a survey*. SIGKDD Explor. Newsl., 2002. 4(2): p. 6-11.
16. Irani, D., et al., *Large Online Social Footprints--An Emerging Threat*, in *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 03*2009, IEEE Computer Society. p. 271-276.
17. Chen, T., et al., *Is more always merrier?: a deep dive into online social footprints*, in *Proceedings of the 2012 ACM workshop on Workshop on online social networks* 2012, ACM: Helsinki, Finland. p. 67-72.
18. Abel, F., et al., *Interweaving public user profiles on the web*, in *Proceedings of the 18th international conference on User Modeling, Adaptation, and Personalization*2010, Springer-Verlag: Big Island, HI. p. 16-27.
19. Minder, P. and A. Bernstein, *Social Network Aggregation Using Face-Recognition*, in *ISWC 2011 Workshop: Social Data on the Web*2011: Bonn, Germany.
20. Rowe, M. and F. Ciravegna, *Disambiguating identity web references using Web 2.0 data and semantics*. Web Semant., 2010. 8(2-3): p. 125-142.
21. Malhotra, A., et al., *Studying User Footprints in Different Online Social Networks*. International Workshop on Cybersecurity of Online Social Network (CSOSN), 2012.

22. Bojars, U., et al., *Towards Semantically-Interlinked Online Communities*. Call for Take up Actions, Joined sub-Projects to AXMEDIS project of the European Commission, 2005: p. 35.
23. Scerri, S., et al., *Knowledge Discovery in distributed Social Web sharing activities*. Making Sense of Microposts (# MSM2012), 2012: p. 26-33.
24. Sauermann, L., L. Van Elst, and A. Dengel, *Pimo-a framework for representing personal information models*. Proceedings of I-Semantics, 2007. 7: p. 270-277.
25. Cortis, K., et al., *Discovering Semantic Equivalence of People behind Online Profiles*. In Proceedings of the Resource Discovery (RED) Workshop, ESWC 2012, 2012.
26. Rowe, M. *Mapping between digital identity ontologies through sism*. in *In Proceedings of the Social Data on the Web Workshop at the International Semantic Web Conference*. 2009.
27. Ninghui, L., L. Tiancheng, and S. Venkatasubramanian. *t-Closeness: Privacy Beyond k-Anonymity and l-Diversity*. in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. 2007.
28. Machanavajjhala, A., et al., *ϵ -diversity: Privacy beyond k-anonymity*. ACM Trans. Knowl. Discov. Data, 2007. 1(1): p. 3.
29. Samarati, P., *Protecting Respondents' Identities in Microdata Release*. IEEE Trans. on Knowl. and Data Eng., 2001. 13(6): p. 1010-1027.
30. Sweeney, L., *k-anonymity: a model for protecting privacy*. Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 2002. 10(5): p. 557-570.
31. Staddon, J., P. Golle, and B. Zimny, *Web-based inference detection*, in *Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium2007*, USENIX Association: Boston, MA. p. 1-16.
32. Chen, Y. and W. Chu, *Database Security Protection Via Inference Detection*, in *Intelligence and Security Informatics*, S. Mehrotra, et al., Editors. 2006, Springer Berlin Heidelberg. p. 452-458.

33. Li, Y.-R., L.-H. Wang, and C.-F. Hong, *Extracting the significant-rare keywords for patent analysis*. Expert Systems with Applications, 2009. 36(3, Part 1): p. 5200-5204.
34. Yip, R.W. and K.N. Levitt. *Data level inference detection in database systems*. in *Computer Security Foundations Workshop, 1998. Proceedings. 11th IEEE*. 1998.
35. Cuppens, F. and A. Gabillon, *Cover story management*. Data & Knowledge Engineering, 2001. 37(2): p. 177-201.
36. Sweeney, L., *Achieving k-anonymity privacy protection using generalization and suppression*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002. 10(05): p. 571-588.
37. Sweeney, L., *k-anonymity: A model for protecting privacy*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002. 10(05): p. 557-570.
38. Geng, L., et al., *Privacy Measures for Free Text Documents: Bridging the Gap between Theory and Practice*, in *Trust, Privacy and Security in Digital Business*, S. Furnell, C. Lambrinoudakis, and G. Pernul, Editors. 2011, Springer Berlin Heidelberg. p. 161-173.
39. Ma, R., X. Meng, and Z. Wang, *Preserving privacy on the searchable internet*, in *Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services* 2011, ACM: Ho Chi Minh City, Vietnam. p. 238-245.
40. Zheleva, E. and L. Getoor, *Privacy in social networks: A survey*, in *Social network data analytics* 2011, Springer. p. 277-306.
41. Schoenmackers, S., O. Etzioni, and D.S. Weld, *Scaling textual inference to the web*, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* 2008, Association for Computational Linguistics: Honolulu, Hawaii. p. 79-88.
42. Richardson, M. and P. Domingos, *Markov logic networks*. Machine Learning, 2006. 62(1-2): p. 107-136.
43. Schoenmackers, S., et al., *Learning first-order Horn clauses from web text*, in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language*

- Processing*2010, Association for Computational Linguistics: Cambridge, Massachusetts. p. 1088-1098.
44. Mori, J., Y. Matsuo, and M. Ishizuka. *Finding user semantics on the web using word co-occurrence information*. in *Proceedings of the International Workshop on Personalization on the Semantic Web (PersWeb05)*. 2005. Citeseer.
 45. Gessiou, E., V. Quang Hieu, and S. Ioannidis. *IRILD: An Information Retrieval Based Method for Information Leak Detection*. in *Computer Network Defense (EC2ND), 2011 Seventh European Conference on*. 2011.
 46. Chow, R., P. Golle, and J. Staddon, *Detecting privacy leaks using corpus-based association rules*, in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*2008, ACM: Las Vegas, Nevada, USA. p. 893-901.
 47. Staddon, J., et al., *A content-driven access control system*, in *Proceedings of the 7th symposium on Identity and trust on the Internet*2008, ACM: Gaithersburg, Maryland, USA. p. 26-35.
 48. Chakaravarthy, V.T., et al., *Efficient techniques for document sanitization*, in *Proceedings of the 17th ACM conference on Information and knowledge management*2008, ACM: Napa Valley, California, USA. p. 843-852.
 49. Nettleton, D. and D. Abril, *Document Sanitization: Measuring Search Engine Information Loss and Risk of Disclosure for the Wikileaks cables*, in *Privacy in Statistical Databases*, J. Domingo-Ferrer and I. Tinnirello, Editors. 2012, Springer Berlin Heidelberg. p. 308-321.
 50. Yang, X. and C. Li, *Secure XML publishing without information leakage in the presence of data inference*, in *Proceedings of the Thirtieth international conference on Very large data bases - Volume 30*2004, VLDB Endowment: Toronto, Canada. p. 96-107.
 51. Thiel, S., A. Schuller, and F. Hermann, *Navigating the Personal Information Sphere*. 1st International Workshop on Model-based Interactive Ubiquitous System 2011, Pisa, Italy, 2011.
 52. Narayanan, A. and V. Shmatikov. *Robust De-anonymization of Large Sparse Datasets*. in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. 2008.

53. Narayanan, A. and V. Shmatikov. *De-anonymizing Social Networks*. in *Security and Privacy, 2009 30th IEEE Symposium on*. 2009.
54. Lam, I.-F., K.-T. Chen, and L.-J. Chen, *Involuntary Information Leakage in Social Network Services*, in *Advances in Information and Computer Security*, K. Matsuura and E. Fujisaki, Editors. 2008, Springer Berlin Heidelberg. p. 167-183.
55. Friedland, G., et al., *Sherlock holmes' evil twin: on the impact of global inference for online privacy*, in *Proceedings of the 2011 workshop on New security paradigms workshop2011*, ACM: Marin County, California, USA. p. 105-114.
56. Lovász, L., *Random walks on graphs: A survey*. Combinatorics, Paul erdos is eighty, 1993. 2(1): p. 1-46.
57. Groza, T., S. Handschuh, and K. Moeller, *The nepomuk project-on the way to the social semantic desktop*. 2007.
58. Dan Brickley, L.M., *Foaf vocabulary specification 0.91*, 2007.
59. Breslin, J.G., et al., *Towards semantically-interlinked online communities*, in *Proceedings of the Second European conference on The Semantic Web: research and Applications2005*, Springer-Verlag: Heraklion, Greece. p. 500-514.
60. Sauermann, L., et al. *Personalization in the EPOS project*. in *Proceedings of the International Workshop on Semantic Web Personalization, Budva, Montenegro*. 2006.
61. Sauermann, L., L.V. Elst, and K. Möller, *Personal information model (PIMO), NEPOMUK Recommendation, v 1.1*. 2009.
62. Martín-Serrano, D., R. Hervás, and J. Bravo, *Telemaco: Context-aware System for Tourism Guiding based on Web 3.0 Technology*.
63. Scerri, S., et al., *digital. me towards an integrated Personal Information Sphere*. Proceedings on the Federated Social Web Summit, 2011: p. W3C.
64. Allan, J., *Introduction to topic detection and tracking*, in *Topic detection and tracking*, A. James, Editor 2002, Kluwer Academic Publishers. p. 1-16.

65. Angajala, P., S. Madria, and M. Linderman, *ECO: Event Detection from Click-through Data via Query Clustering*, in *On the Move to Meaningful Internet Systems: OTM 2012*, R. Meersman, et al., Editors. 2012, Springer Berlin Heidelberg. p. 323-341.
66. Yang, Y., et al., *Learning approaches for detecting and tracking news events*. IEEE Intelligent Systems, 1999. 14(4): p. 32-43.
67. Nallapati, R., et al., *Event threading within news topics*, in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*2004, ACM: Washington, D.C., USA. p. 446-453.
68. Qin, Y., et al. *Feature-Rich Segment-Based News Event Detection on Twitter*. in *International Joint Conference on Natural Language Processing*. 2013.
69. Xu, Z., et al., *Knowle: A semantic link network based system for organizing large scale online news events*. Future Generation Computer Systems, 2014(0).
70. Makkonen, J., H. Ahonen-Myka, and M. Salmenkivi. *Applying semantic classes in event detection and tracking*. in *Proceedings of International Conference on Natural Language Processing (ICON 2002)*. 2002. Citeseer.
71. Wang, Y., H. Sundaram, and L. Xie, *Social event detection with interaction graph modeling*, in *Proceedings of the 20th ACM international conference on Multimedia*2012, ACM: Nara, Japan. p. 865-868.
72. Zaharieva, M., M. Zeppelzauer, and C. Breiteneder, *Automated social event detection in large photo collections*, in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*2013, ACM: Dallas, Texas, USA. p. 167-174.
73. Liu, X., et al., *Using social media to identify events*, in *Proceedings of the 3rd ACM SIGMM international workshop on Social media*2011, ACM: Scottsdale, Arizona, USA. p. 3-8.
74. Rattenbury, T., N. Good, and M. Naaman, *Towards automatic extraction of event and place semantics from flickr tags*, in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*2007, ACM: Amsterdam, The Netherlands. p. 103-110.

75. Mahata, D. and N. Agarwal, *Learning from the crowd: an evolutionary mutual reinforcement model for analyzing events*, in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*2013, ACM: Niagara, Ontario, Canada. p. 474-478.
76. Allen, J.F. and G. Ferguson, *Actions and Events in Interval Temporal Logic*, 1994, University of Rochester.
77. Valkanas, G. and D. Gunopulos, *How the live web feels about events*, in *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management*2013, ACM: San Francisco, California, USA. p. 639-648.
78. Paniagua, J., et al., *Social events and social ties*, in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*2013, ACM: Dallas, Texas, USA. p. 143-150.
79. Sakaki, T., M. Okazaki, and Y. Matsuo, *Earthquake shakes Twitter users: real-time event detection by social sensors*, in *Proceedings of the 19th international conference on World wide web*2010, ACM: Raleigh, North Carolina, USA. p. 851-860.
80. Gupta, M., P. Zhao, and J. Han, *Evaluating Event Credibility on Twitter*, in *SDM2012*, SIAM / Omnipress. p. 153-164.
81. Sarma, A.D., A. Jain, and C. Yu, *Dynamic relationship and event discovery*, in *Proceedings of the fourth ACM international conference on Web search and data mining*2011, ACM: Hong Kong, China. p. 207-216.
82. Sheba, S., B. Ramadoss, and S.R. Balasundaram, *Event Detection Refinement Using External Tags for Flickr Collections*, in *Intelligent Computing, Networking, and Informatics*, D.P. Mohapatra and S. Patnaik, Editors. 2014, Springer India. p. 369-375.
83. Tankoyeu, I., et al., *Event detection and scene attraction by very simple contextual cues*, in *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*2011, ACM: Scottsdale, Arizona, USA. p. 1-6.
84. Westermann, U. and R. Jain, *Toward a Common Event Model for Multimedia Applications*. MultiMedia, IEEE, 2007. 14(1): p. 19-29.

85. Abdelhaq, H., C. Sengstock, and M. Gertz, *EvenTweet: online localized event detection from twitter*. Proc. VLDB Endow., 2013. 6(12): p. 1326-1329.
86. Bauer, A. and C. Wolff, *Event based classification of Web 2.0 text streams*. arXiv preprint arXiv:1204.3362, 2012.
87. Becker, H., M. Naaman, and L. Gravano, *Beyond Trending Topics: Real-World Event Identification on Twitter*. ICWSM, 2011. 11: p. 438-441.
88. Kumar, S., et al., *From Tweets to Events: Exploring a Scalable Solution for Twitter Streams*. arXiv preprint arXiv:1405.1392, 2014.
89. Liu, X., et al. *Exacting Social Events for Tweets Using a Factor Graph*. in AAAI. 2012.
90. Marcus, A., et al., *Twitinfo: aggregating and visualizing microblogs for event exploration*, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*2011, ACM: Vancouver, BC, Canada. p. 227-236.
91. Bao, B.-K., et al., *Social event detection with robust high-order co-clustering*, in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*2013, ACM: Dallas, Texas, USA. p. 135-142.
92. Ballan, L., et al., *Event detection and recognition for semantic annotation of video*. Multimedia Tools and Applications, 2011. 51(1): p. 279-302.
93. Gkonela, C. and K. Chorianopoulos, *VideoSkip: event detection in social web videos with an implicit user heuristic*. Multimedia Tools and Applications, 2014. 69(2): p. 383-396.
94. Mertens, R., et al., *Acoustic super models for large scale video event detection*, in *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*2011, ACM: Scottsdale, Arizona, USA. p. 19-24.
95. Steiner, T., et al., *Crowdsourcing event detection in youtube videos*. Proc. Detection, Representation, and Exploitation of Events in the Semantic Web, 2011: p. 58-67.
96. Yan, W., et al., *A comprehensive study of visual event computing*. Multimedia Tools and Applications, 2011. 55(3): p. 443-481.

97. Scherp, A. and V. Mezaris, *Survey on modeling and indexing events in multimedia*. Multimedia Tools and Applications, 2014. 70(1): p. 7-23.
98. Chen, L. and A. Roy, *Event detection from flickr data through wavelet-based spatial analysis*, in *Proceedings of the 18th ACM conference on Information and knowledge management*2009, ACM: Hong Kong, China. p. 523-532.
99. Firan, C.S., et al., *Bringing order to your photos: event-driven classification of flickr images based on social knowledge*, in *Proceedings of the 19th ACM international conference on Information and knowledge management*2010, ACM: Toronto, ON, Canada. p. 189-198.
100. Papadopoulos, S., et al., *Cluster-Based Landmark and Event Detection for Tagged Photo Collections*. MultiMedia, IEEE, 2011. 18(1): p. 52-63.
101. Rafatirad, S., R. Jain, and K. Laskey. *Context-Based Event Ontology Extension in Multimedia Applications*. in *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on*. 2013.
102. Ruocco, M. and H. Ramampiaro, *A scalable algorithm for extraction and clustering of event-related pictures*. Multimedia Tools and Applications, 2014. 70(1): p. 55-88.
103. Brenner, M. and E. Izquierdo, *Social event detection and retrieval in collaborative photo collections*, in *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*2012, ACM: Hong Kong, China. p. 1-8.
104. Dao, M.-S., D.-T. Dang-Nguyen, and F.G. De Natale, *Robust event discovery from photo collections using Signature Image Bases (SIBs)*. Multimedia Tools and Applications, 2012: p. 1-29.
105. Becker, H., M. Naaman, and L. Gravano, *Learning similarity metrics for event identification in social media*, in *Proceedings of the third ACM international conference on Web search and data mining*2010, ACM: New York, New York, USA. p. 291-300.
106. Natarajan, P. and R. Nevatia, *EDF: A framework for Semantic Annotation of Video*, in *Proceedings of the Tenth IEEE International Conference on Computer Vision Workshops*2005, IEEE Computer Society. p. 1876.

107. Pohl, D., A. Bouchachia, and H. Hellwagner, *Automatic sub-event detection in emergency management using social media*, in *Proceedings of the 21st international conference companion on World Wide Web2012*, ACM: Lyon, France. p. 683-686.
108. Rafatirad, S., A. Gupta, and R. Jain, *Event composition operators: ECO*, in *Proceedings of the 1st ACM international workshop on Events in multimedia2009*, ACM: Beijing, China. p. 65-72.
109. Wang, X.-j., et al., *Eventory -- An Event Based Media Repository*, in *Proceedings of the International Conference on Semantic Computing2007*, IEEE Computer Society. p. 95-104.
110. Kleinberg, J., *Bursty and hierarchical structure in streams*. Data Mining and Knowledge Discovery, 2003. 7(4): p. 373-397.
111. Güting, R.H., *An introduction to spatial database systems*. The VLDB Journal—The International Journal on Very Large Data Bases, 1994. 3(4): p. 357-399.
112. Peuquet, D.J. and Z. Ci-Xiang, *An algorithm to determine the directional relationship between arbitrarily-shaped polygons in the plane*. Pattern recognition, 1987. 20(1): p. 65-74.
113. Anwar, F., et al. *An Efficient Event Definition Framework for Retail Sector Surveillance Systems*. in *MMEDIA 2014, The Sixth International Conferences on Advances in Multimedia*. 2014.
114. Li, S. and H. Wang, *RCC8 binary constraint network can be consistently extended*. Artificial Intelligence, 2006. 170(1): p. 1-18.
115. Rafatirad, S., *Context-based event ontology extension in multimedia applications*, 2012, California State University at Long Beach. p. 153.
116. Gupta, A. and R. Jain, *Managing event information: Modeling, retrieval, and applications*. Synthesis Lectures on Data Management, 2011. 3(4): p. 1-141.
117. Abhik, D. and D. Toshniwal, *Sub-event detection during natural hazards using features of social media data*, in *Proceedings of the 22nd international conference on World Wide Web companion2013*, International World Wide Web Conferences Steering Committee: Rio de Janeiro, Brazil. p. 783-788.

118. Feng, X., et al., *Real-Time Event Detection Based on Geo Extraction and Temporal Analysis*, in *Advanced Data Mining and Applications*, X. Luo, J. Yu, and Z. Li, Editors. 2014, Springer International Publishing. p. 137-150.
119. Klein, B., et al., *Detection and Extracting of Emergency Knowledge from Twitter Streams*, in *Ubiquitous Computing and Ambient Intelligence*, J. Bravo, D. López-de-Ipiña, and F. Moya, Editors. 2012, Springer Berlin Heidelberg. p. 462-469.
120. Panteras, G., et al., *Triangulating Social Multimedia Content for Event Localization using Flickr and Twitter*. *Transactions in GIS*, 2014: p. n/a-n/a.
121. Vieweg, S., et al., *Microblogging during two natural hazards events: what twitter may contribute to situational awareness*, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*2010, ACM: Atlanta, Georgia, USA. p. 1079-1088.
122. Zin, T.T., et al. *Knowledge based Social Network Applications to Disaster Event Analysis*. in *Proceedings of the International MultiConference of Engineers and Computer Scientists*. 2013.
123. Pohl, D., A. Bouchachia, and H. Hellwagner. *Online Processing of Social Media Data for Emergency Management*. in *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*. 2013.
124. Bouchachia, A. and C. Vanaret. *Incremental Learning Based on Growing Gaussian Mixture Models*. in *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*. 2011.
125. Dao, M.-S., et al., *Jointly exploiting visual and non-visual information for event-related social media retrieval*, in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*2013, ACM: Dallas, Texas, USA. p. 159-166.
126. Dwork, C., *Differential privacy*, in *Encyclopedia of Cryptography and Security*2011, Springer. p. 338-340.
127. Huang, J., S. Ertekin, and C.L. Giles, *Efficient name disambiguation for large-scale databases*, in *Knowledge Discovery in Databases: PKDD 2006*2006, Springer. p. 536-544.

128. Press, W.H., *Numerical recipes 3rd edition: The art of scientific computing* 2007: Cambridge university press.
129. Bouna, B.A., R. Chbeir, and A. Gabillon. *The image protector-a flexible security rule specification toolkit*. in *Security and Cryptography (SECRYPT), 2011 Proceedings of the International Conference on*. 2011. IEEE.
130. Fan, J., et al. *A novel approach for privacy-preserving video sharing*. in *Proceedings of the 14th ACM international conference on Information and knowledge management*. 2005. ACM.
131. Boyle, M., C. Edwards, and S. Greenberg. *The effects of filtered video on awareness and privacy*. in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 2000. ACM.
132. Chun, B.T., Y. Bae, and T.-Y. Kim. *A method for original image recovery for caption areas in video*. in *Systems, Man, and Cybernetics, 1999. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on*. 1999. IEEE.
133. Criminisi, A., P. Pérez, and K. Toyama, *Region filling and object removal by exemplar-based image inpainting*. Image Processing, IEEE Transactions on, 2004. 13(9): p. 1200-1212.
134. Komodakis, N. *Image completion using global optimization*. in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. 2006. IEEE.
135. Patwardhan, K., G. Sapiro, and M. Bertalmío, *Video inpainting under constrained camera motion*. Image Processing, IEEE Transactions on, 2007. 16(2): p. 545-553.
136. Wang, L., et al. *Stereoscopic inpainting: Joint color and depth completion from stereo images*. in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. 2008. IEEE.
137. Wexler, Y., E. Shechtman, and M. Irani, *Space-time completion of video*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2007. 29(3): p. 463-476.
138. Chow, R., P. Golle, and J. Staddon. *Detecting privacy leaks using corpus-based association rules*. in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008. ACM.

139. LeFevre, K., D.J. DeWitt, and R. Ramakrishnan. *Mondrian multidimensional k-anonymity*. in *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*. 2006. IEEE.
140. Xiao, X. and Y. Tao. *Anatomy: Simple and effective privacy preservation*. in *Proceedings of the 32nd international conference on Very large data bases*. 2006. VLDB Endowment.
141. Martin, N.F. and J.W. England, *Mathematical theory of entropy*. Vol. 12. 2011: Cambridge university press.
142. Saini, M., et al., *W3-privacy: understanding what, when, and where inference channels in multi-camera surveillance video*. *Multimedia Tools Appl.*, 2014. 68(1): p. 135-158.
143. Hirota, M., et al., *A robust clustering method for missing metadata in image search results*. *Journal of Information Processing*, 2012. 20(3): p. 537-547.
144. Ku, W., M. Kankanhalli, and J.-H. Lim, *Metadata Management, Reuse, Inference and Propagation in a Collection-Oriented Metadata Framework for Digital Images*, in *Advances in Multimedia Modeling*, T.-J. Cham, et al., Editors. 2006, Springer Berlin Heidelberg. p. 145-154.
145. Raad, E., R. Chbeir, and A. Dipanda, *Discovering relationship types between users using profiles and shared photos in a social network*. *Multimedia Tools and Applications*, 2013. 64(1): p. 141-170.
146. Backstrom, L. and J. Kleinberg, *Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on facebook*, in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing2014*, ACM: Baltimore, Maryland, USA. p. 831-841.
147. Wu, D., N. Mamoulis, and J. Shi, *Clustering in Geo-Social Networks*. *IEEE Data Engineering Bulletin*, 2015.
148. Rossi, L. and M. Musolesi, *It's the way you check-in: identifying users in location-based social networks*, in *Proceedings of the second ACM conference on Online social networks2014*, ACM: Dublin, Ireland. p. 215-226.

149. Yan, R., et al., *Evolutionary timeline summarization: a balanced optimization framework via iterative substitution*, in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* 2011, ACM: Beijing, China. p. 745-754.
150. Yang, C.C. and X. Shi. *Discovering event evolution graphs from newswires*. in *Proceedings of the 15th international conference on World Wide Web*. 2006. ACM.
151. Hobona, G., P. James, and D. Fairbairn, *Multidimensional visualisation of degrees of relevance of geographic data*. *International Journal of Geographical Information Science*, 2006. 20(05): p. 469-490.
152. Riise, S. and D. Patel, *System and method for creating minimum bounding rectangles for use in a geo-coding system*, 2011, Google Patents.
153. Allen, J.F. and G. Ferguson, *Actions and events in interval temporal logic*. *Journal of logic and computation*, 1994. 4(5): p. 531-579.
154. Taddesse, F.G., et al., *Semantic-based merging of RSS items*. *World Wide Web*, 2010. 13(1-2): p. 169-207.
155. Berkhin, P., *A Survey of Clustering Data Mining Techniques*, in *Grouping Multidimensional Data*, J. Kogan, C. Nicholas, and M. Teboulle, Editors. 2006, Springer Berlin Heidelberg. p. 25-71.
156. Hubert, L.J. and J.R. Levin, *A general statistical framework for assessing categorical clustering in free recall*. *Psychological Bulletin*, 1976. 83(6): p. 1072.
157. Miller, Z., et al., *Twitter spammer detection using data stream clustering*. *Inf. Sci.*, 2014. 260: p. 64-73.
158. Reuter, T., et al., *ReSEED: social event dEtection dataset*, in *Proceedings of the 5th ACM Multimedia Systems Conference* 2014, ACM: Singapore, Singapore. p. 35-40.
159. Bouma, G., *Normalized (pointwise) mutual information in collocation extraction*. *Proceedings of GSCL*, 2009: p. 31-40.
160. Strehl, A. and J. Ghosh, *Relationship-based clustering and visualization for high-dimensional data mining*. *INFORMS Journal on Computing*, 2003. 15(2): p. 208-230.

161. Pluim, J.P., J.A. Maintz, and M.A. Viergever. *Image registration by maximization of combined mutual information and gradient information*. in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2000*. 2000. Springer.
162. Shiga, M., I. Takigawa, and H. Mamitsuka, *Annotating gene function by combining expression data with a modular gene network*. *Bioinformatics*, 2007. 23(13): p. i468-i478.
163. Zhong, S. and J. Ghosh, *A unified framework for model-based clustering*. *The Journal of Machine Learning Research*, 2003. 4: p. 1001-1037.
164. Zhong, S. and J. Ghosh, *Generative model-based document clustering: a comparative study*. *Knowledge and Information Systems*, 2005. 8(3): p. 374-384.
165. Papadopoulos, S., et al. *The 2012 social event detection dataset*. in *Proceedings of the 4th ACM Multimedia Systems Conference*. 2013. ACM.
166. Petkos, G., et al. *Social event detection at mediaeval: a three-year retrospect of tasks and results*. in *Proc. ACM ICMR 2014 Workshop on Social Events in Web Multimedia (SEWM)*. 2014.
167. Rafailidis, D., et al. *A Data-Driven Approach for Social Event Detection*. in *MediaEval*. 2013.
168. Manchon Vizuite, D., I. Gris-Sarabia, and X. Giró Nieto. *UPC at MediaEval 2014 social event detection task*. 2014. CEUR Workshop Proceedings.
169. Nguyen, T.-V., et al. *Event Clustering and Classification from Social Media: Watershed-based and Kernel Methods*. in *MediaEval*. 2013.
170. Schinas, M., et al., *CERTH @ MediaEval 2013 Social Event Detection Task*. Working Notes Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, October 18-19, CEUR-WS.org, ISSN 1613-0073, 2013.
171. Wistuba, M. and L. Schmidt-Thieme. *Supervised Clustering of Social Media Streams*. in *MediaEval*. 2013.
172. Tserpes, K., M. Kardara, and T. Varvarigou, *A similarity-based Chinese Restaurant Process for Social Event Detection*.

173. Samangooei, S., et al., *Social event detection via sparse multi-modal feature selection and incremental density based clustering*. 2013.
174. Zeppelzauer, M., M. Zaharieva, and M. Del Fabro. *Unsupervised Clustering of Social Events*. in *MediaEval*. 2013.
175. Sutanto, T. and R. Nayak. *Admrg@ mediaeval 2013 social event detection*. in *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*. 2013. CEUR-WS. org.
176. Itika Gupta, K.G. and K. Chandramouli, *VIT@ MediaEval 2013 Social Event Detection Task: Semantic Structuring of Complementary Information for Clustering Events*.
177. Brenner, M. and E. Izquierdo, *MediaEval 2013: Social Event Detection, Retrieval and Classification in Collaborative Photo Collections*. MediaEval, 2013. 1043.
178. Yan, W., et al., *A comprehensive study of visual event computing*. Multimedia Tools Appl., 2011. 55(3): p. 443-481.
179. Hung, C.-C., et al. *Tag-based user profiling for social media recommendation*. in *Workshop on Intelligent Techniques for Web Personalization & Recommender Systems at AAAI*. 2008.
180. Li, R., et al. *Towards social user profiling: unified and discriminative influence model for inferring home locations*. in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012. ACM.
181. Zeng, D., et al., *Social media analytics and intelligence*. Intelligent Systems, IEEE, 2010. 25(6): p. 13-16.
182. Karampelas, P., *Visual Methods and Tools for Social Network Analysis*, in *Encyclopedia of Social Network Analysis and Mining* 2014, Springer. p. 2314-2327.
183. Wasserman, S. and K. Faust, *Social network analysis: Methods and applications*. Vol. 8. 1994: Cambridge university press.

184. Rabbath, M. and S. Boll, *Detecting Multimedia Contents of Social Events in Social Networks*, in *Social Media Retrieval*, N. Ramzan, et al., Editors. 2013, Springer London. p. 87-111.
185. Abowd, G.D., et al., *Towards a Better Understanding of Context and Context-Awareness*, in *Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing* 1999, Springer-Verlag: Karlsruhe, Germany. p. 304-307.
186. Hernández-Gracidas, C. and L.E. Sucar, *Markov random fields and spatial information to improve automatic image annotation*, in *Advances in Image and Video Technology* 2007, Springer. p. 879-892.
187. Lienhardt, P., *Topological models for boundary representation: a comparison with n-dimensional generalized maps*. Computer-aided design, 1991. 23(1): p. 59-82.
188. Egenhofer, M.J., J. Sharma, and D.M. Mark. *A critical comparison of the 4-intersection and 9-intersection models for spatial relations: formal analysis*. in *AUTOCARTO-CONFERENCE*-. 1993. ASPRS AMERICAN SOCIETY FOR PHOTOGRAMMETRY AND.
189. Clementini, E., P. Di Felice, and P. van Oosterom. *A small set of formal topological relationships suitable for end-user interaction*. in *Advances in Spatial Databases*. 1993. Springer.
190. Patil, A., J. Liu, and J. Gao. *Predicting group stability in online social networks*. in *Proceedings of the 22nd international conference on World Wide Web*. 2013. International World Wide Web Conferences Steering Committee.
191. Raad, E., *Découverte des relations dans les réseaux sociaux*, 2011, Université de Bourgogne.
192. Xiang, R., J. Neville, and M. Rogati. *Modeling relationship strength in online social networks*. in *Proceedings of the 19th international conference on World wide web*. 2010. ACM.
193. Zhao, X., et al., *Relationship strength estimation for online social networks with the study on Facebook*. Neurocomputing, 2012. 95: p. 89-97.

194. Liu, X., B. Huet, and R. Troncy. *EURECOM@ MediaEval 2011 Social Event Detection Task*. in *MediaEval*. 2011.
195. Wang, Z.J., Z. Han, and K.J.R. Liu, *A MIMO-OFDM channel estimation approach using time of arrivals*. *Trans. Wireless. Comm.*, 2005. 4(3): p. 1207-1213.
196. Mihalcea, R., P. Tarau, and E. Figa, *PageRank on semantic networks, with application to word sense disambiguation*, in *Proceedings of the 20th international conference on Computational Linguistics 2004*, Association for Computational Linguistics: Geneva, Switzerland. p. 1126.
197. Nguyen, D.T. and J.E. Jung, *Privacy-preserving discovery of topic-based events from social sensor signals: An experimental study on twitter*. *The Scientific World Journal*, 2014. 2014.
198. Papadopoulos, S., D. Corney, and L.M. Aiello. *SNOW 2014 Data Challenge: Assessing the Performance of News Topic Detection Methods in Social Media*. in *SNOW-DC@ WWW*. 2014.
199. Unankard, S., X. Li, and M. Sharaf, *Location-Based Emerging Event Detection in Social Networks*, in *Web Technologies and Applications*, Y. Ishikawa, et al., Editors. 2013, Springer Berlin Heidelberg. p. 280-291.
200. Yee, C., N. Keane, and L. Zhou. *Modeling and Characterizing Social Media Topics Using the Gamma Distribution*. in *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*. 2015.
201. Zhu, T. and J. Yu, *A Prerecognition Model for Hot Topic Discovery Based on Microblogging Data*. *The Scientific World Journal*, 2014. 2014.
202. Hopkins, M. *Strategies for determining causes of events*. in *AAAI/IAAI*. 2002.
203. Mueller, R.M. *A Meta-model for Inferring Inter-theory Relationships of Causal Theories*. in *System Sciences (HICSS), 2015 48th Hawaii International Conference on*. 2015.
204. Luo, J., et al., *Geotagging in multimedia and computer vision—a survey*. *Multimedia Tools and Applications*, 2011. 51(1): p. 187-211.

