Analyse exploratoire de flots de liens pour la détection d'événements

Sébastien Heymann

Thèse encadrée par M. Latapy et C. Magnien pendant 2 ans et par B. Le Grand la 3e année.

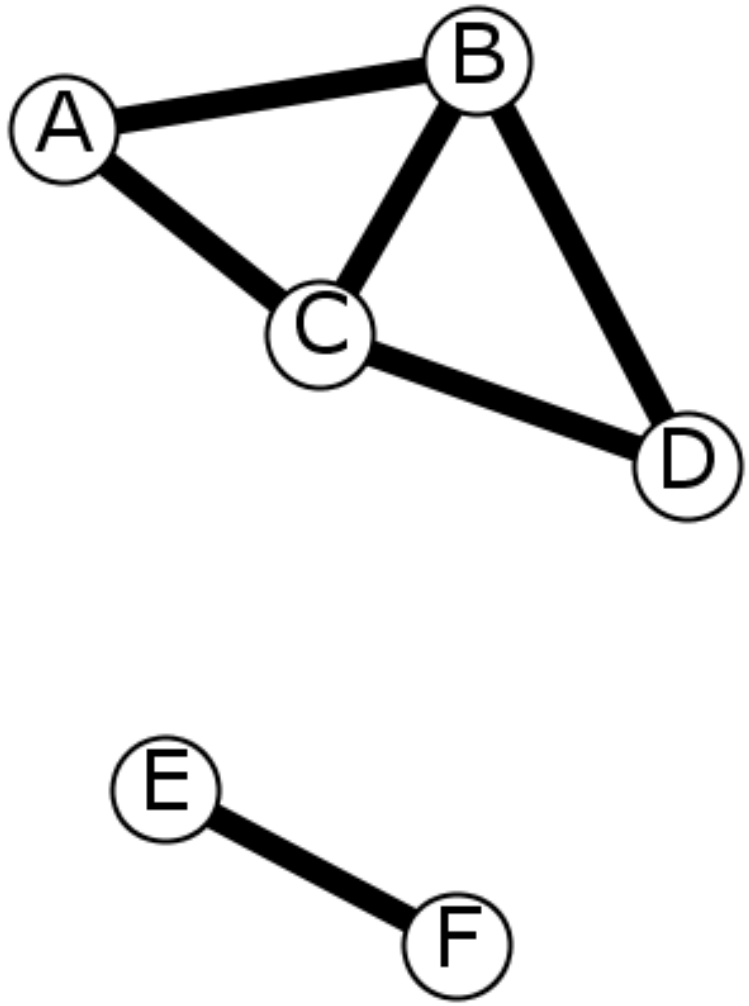
Flot de liens

= trace d'interactions ordonnées chronologiquement entre différentes paires d'entités.

1 : (A) -[interagit avec]- (B)
2 : (C) -[interagit avec]- (D)
3 : (B) -[interagit avec]- (C)
4 : (C) -[interagit avec]- (A)
5 : (B) -[interagit avec]- (D)
6 : (E) -[interagit avec]- (F)
...

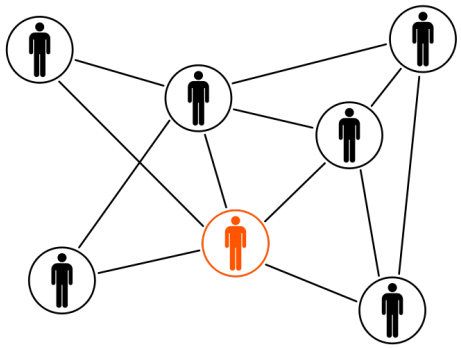
Graphe de flot de liens

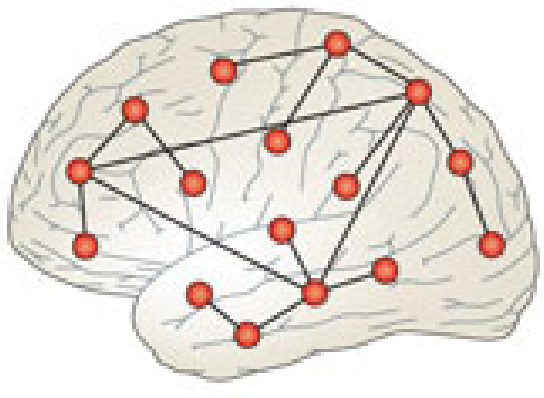
1 : (A) -[interagit avec]- (B)
2 : (C) -[interagit avec]- (D)
3 : (B) -[interagit avec]- (C)
4 : (C) -[interagit avec]- (A)
5 : (B) -[interagit avec]- (D)
6 : (E) -[interagit avec]- (F)
...

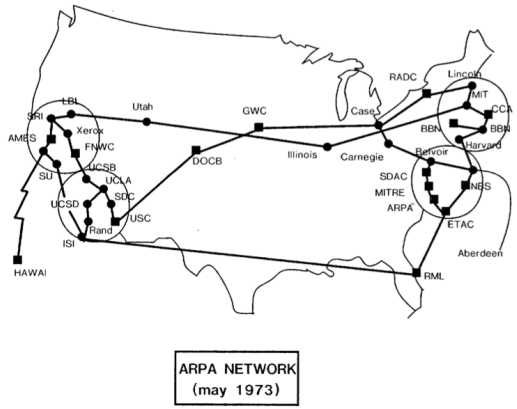


Représentation du monde

Les graphes représentent des mesures de **systèmes complexes** réels ou simulés.

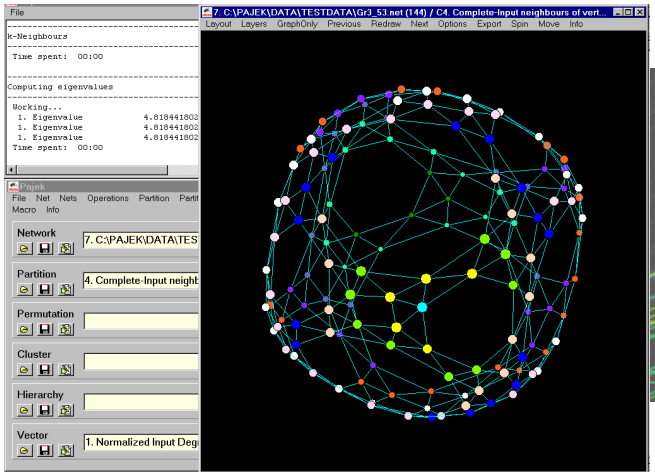


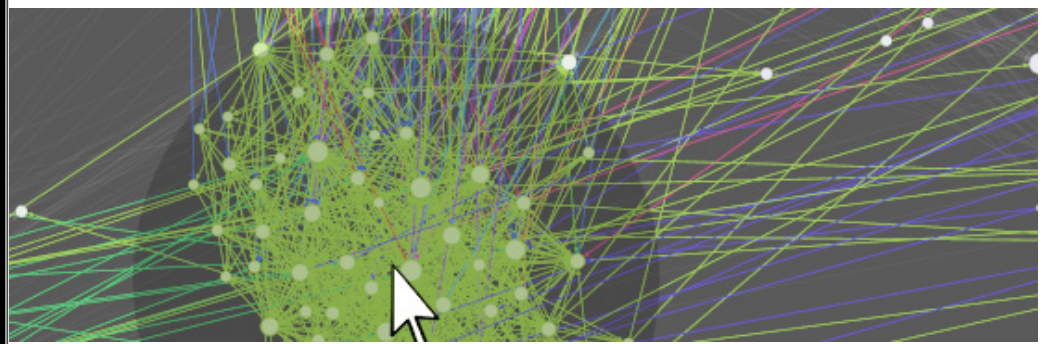




mesure ? analyse ? visualisation ? modélisation ? prédiction ?

Analyse exploratoire de graphes





1 voir le graphe

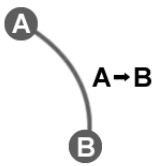
ex : diagramme noeuds-liens

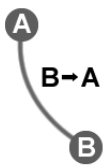
2 interagir en temps réel

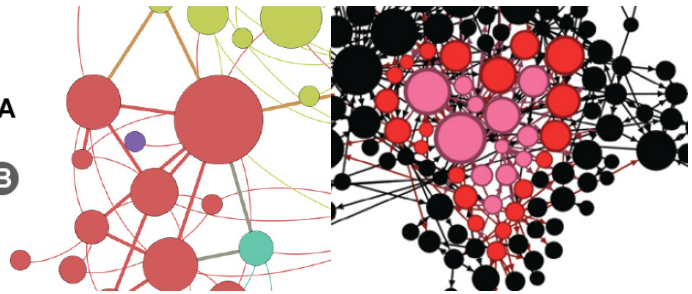
ex: Gephi (2008)
grouper, filtrer, calculer des métriques...

3 construire un langage visuel

couleurs, volumes, formes...







Publications : [chap. Gephi Comp.Net 14]

Pourquoi visualiser des données ?

« The greatest value of a picture
is when it forces us to notice
what we never expected to see. »

John Tukey (1962)

Une visualisation dynamique de flot de liens

André Panisson - The Egyptian Revolution on Twitter, made with Gephi

Problématique

Concepts à définir et formaliser

- Dynamique régulière
- Événement significatif

Contraintes

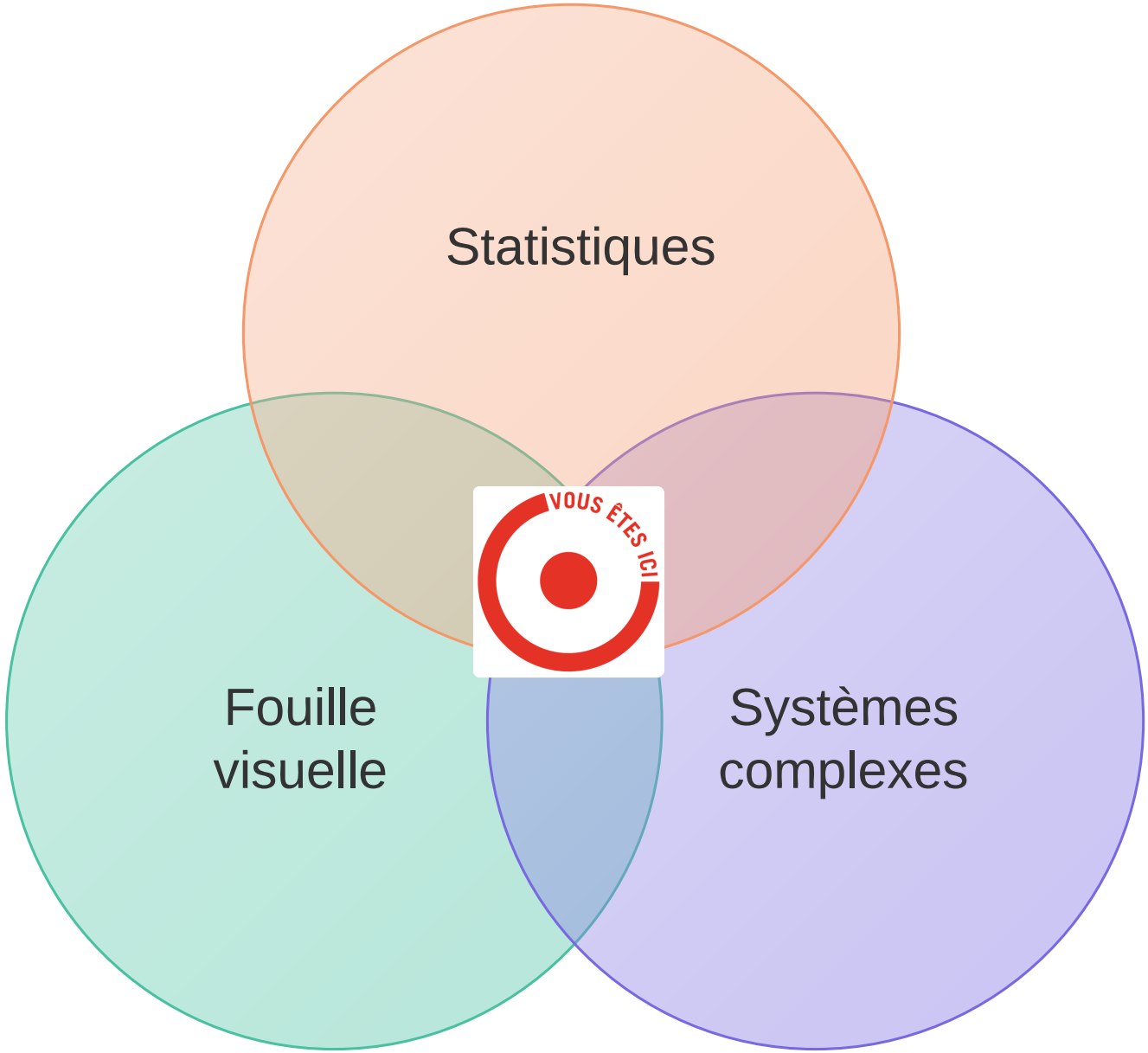
- Démarche exploratoire sans connaissance a priori
- Passage à l'échelle

Objectifs

- Quand scruter des *événements* ?
- Qui est impliqué dans chaque événement ?

Comment détecter des événements ?

- Événement = anomalie temporelle.
- Pas de définition claire et universelle.
- Dépend des cas et des hypothèses de départ.
- Anomalies : identifier les valeurs qui dévient remarquablement des autres (Grubbs, 1969).


Statistiques
VOUS ÊTES ICI
Fouille visuelle
Systèmes complexes

Contributions

1. **Détection automatique d'anomalies.**
   - **nouvelle méthode statistique : Outskewer**
   - **validation expérimentale**
2. Détection automatique d'événements.
   - notion de fenêtre temporelle
   - unité de temps : intrinsèque vs extrinsèque
   - largeur de fenêtre
3. Cadre exploratoire et étude de cas.
   - méthodologie unifiée : statistiques et visualisation
   - prototype
   - application à Github.com

Détection d'anomalies : approaches classiques

Hypothèse : données $\sim$ distribution normale.

Distance valeurs réelles / valeurs théoriques.

Coefficient d'asymétrie

$$\gamma = \frac{n}{(n-1)(n-2)} \sum_{x \in X} \left(\frac{x - \text{moyenne}}{\text{écart type}} \right)^3$$

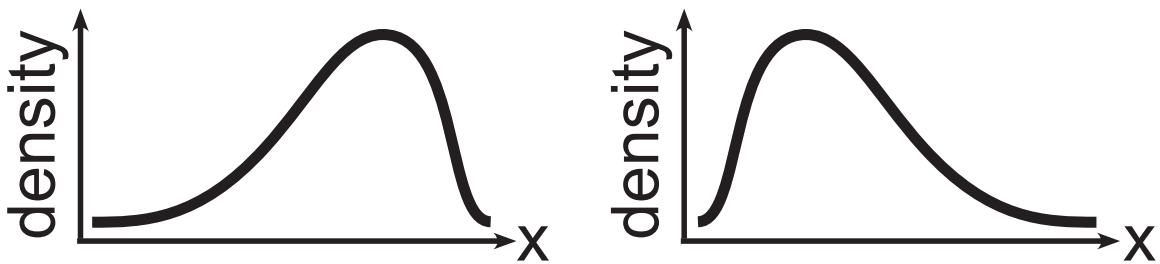


Figure: $\gamma < 0$ $\gamma > 0$

Exemple de distributions asymétriques.

Coefficient d'asymétrie

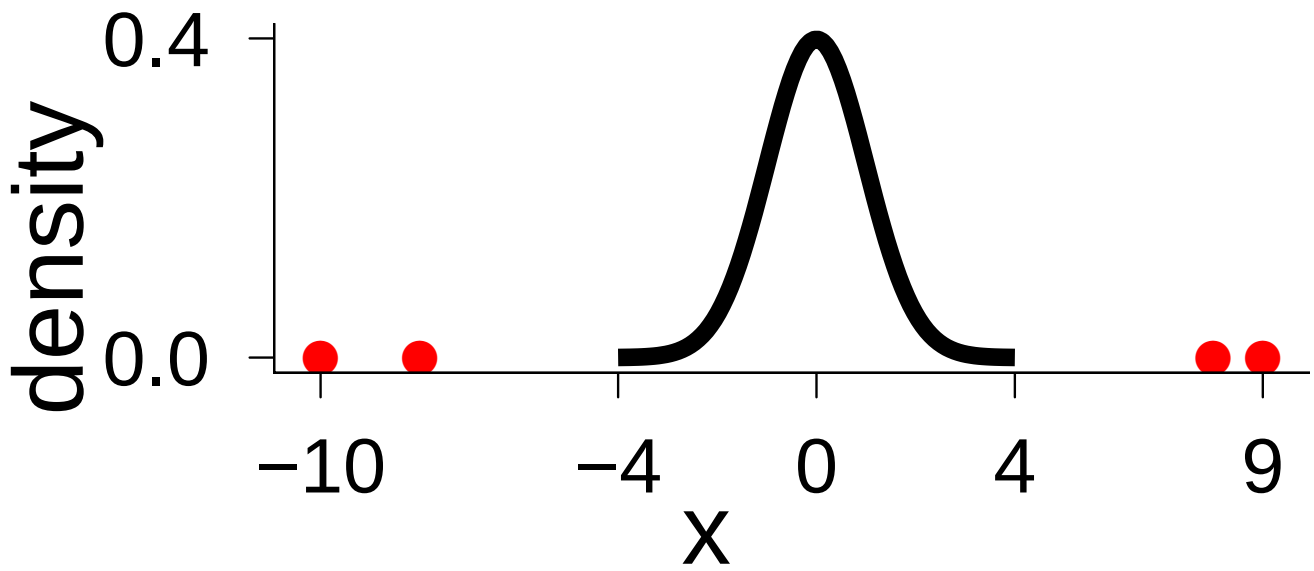


Figure: Exemple de distribution normale avec valeurs extrêmes (rouge).

Sensible aux valeurs extrêmes (min/max) loin de la moyenne !

Postulat de *Outskewer*

Notre définition

Anomalie = valeur extrême qui rend la distribution asymétrique.

Notre méthode

Signature d'asymétrie = évolution du coefficient d'asymétrie quand on retire les valeurs extrêmes une à une.

Implication (distribution homogène)

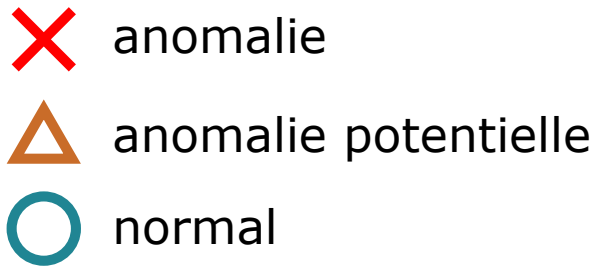
Si présence d'anomalies : l'asymétrie devrait tendre vers 0.

Implication (distribution hétérogène)

Si le retrait d'un grand nombre de valeurs ne réduit pas l'asymétrie, alors la distribution est hétérogène, donc il n'existe pas d'anomalie selon notre définition.

Détection automatique d'anomalies

Analyse de la signature d'asymétrie avec seuillage adaptatif.
Outskewer classe chaque valeur comme :

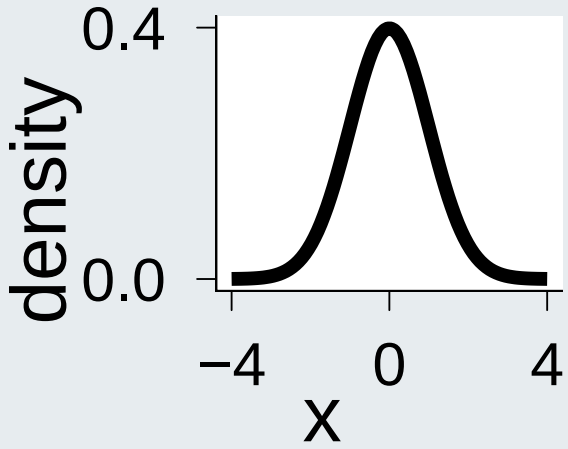


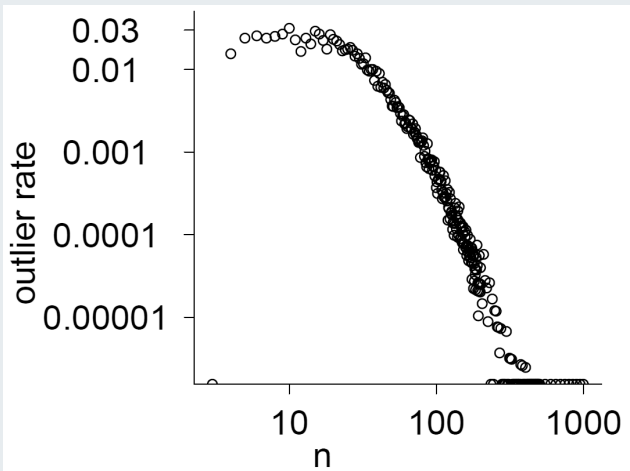
ou 'inconnu' pour les distributions de valeurs hétérogènes.

$X = \{-3, -2, -1, 0, 1, 2, 3, 7\}$

Pertinence pour les graphes réels ?

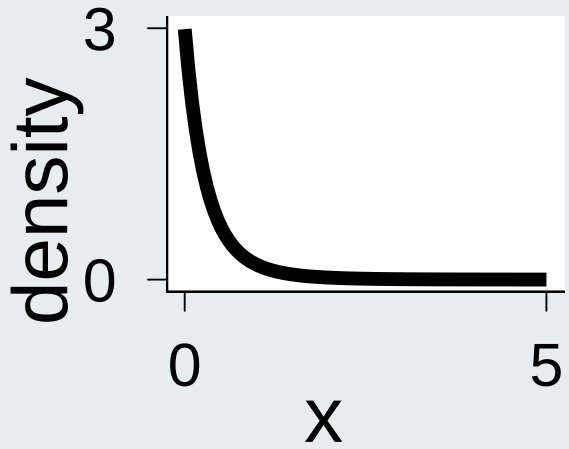
Loi normale (cas homogène)

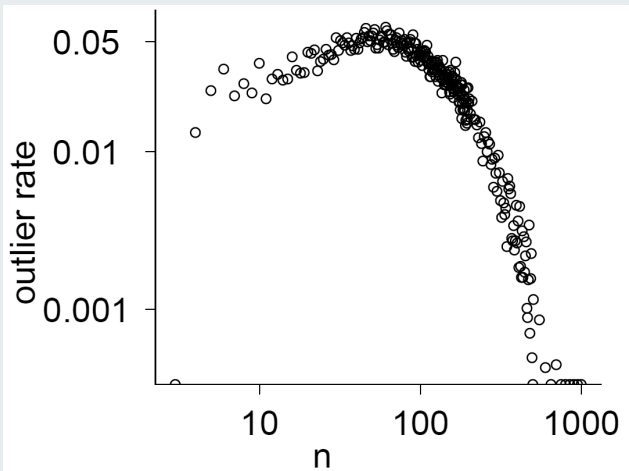




0.01% d'anomalies pour $n = 100$

Loi de puissance (cas hétérogène)





0.01% d'anomalies pour $n = 1000$

Conclusions sur *Outskewer*

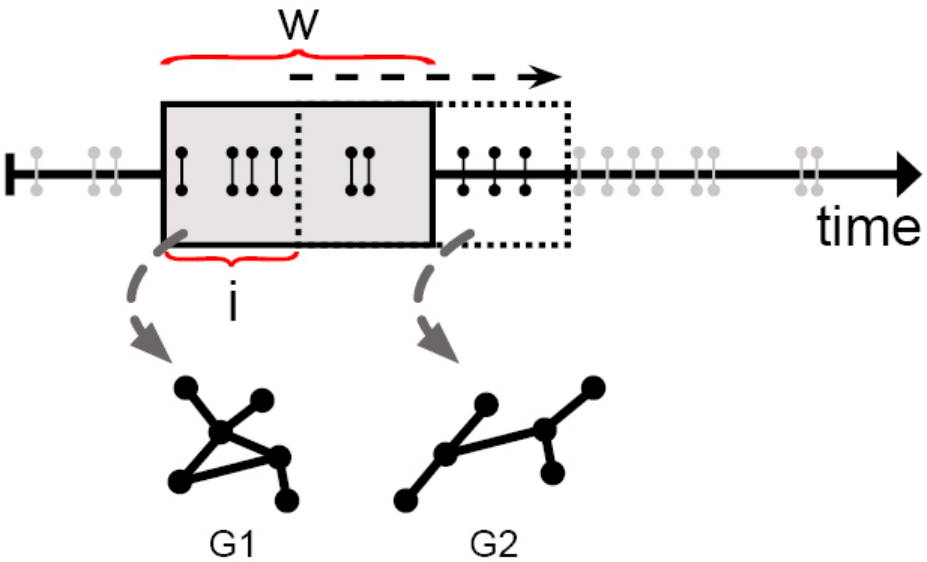
- Méthode de détection automatique d'anomalies.
- Pas de connaissance *a priori* sur les données.
- Pertinent dans différents cas statistiques.

Publication : [ASONAM 12]

Contributions

1. Détection automatique d'anomalies.
   - nouvelle méthode statistique : *Outskewer*
   - validation expérimentale
2. **Détection automatique d'événements.**
   - **notion de fenêtre temporelle**
   - **unité de temps : intrinsèque vs extrinsèque**
   - **largeur de fenêtre**
3. Cadre exploratoire et étude de cas.
   - méthodologie unifiée : statistiques et visualisation
   - prototype
   - application à Github.com

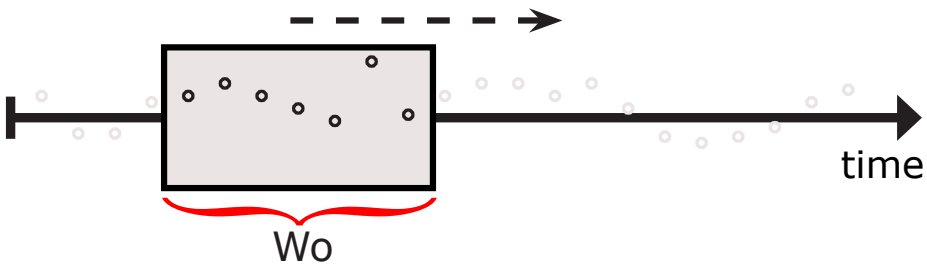
Détection d'événements



1. Mesure

a) fenêtre temporelle de taille w **sur un flot de liens**.
b) graphe de chaque fenêtre à intervalle i.
c) propriété calculée pour chaque graphe : série temporelle.

Détection d'événements



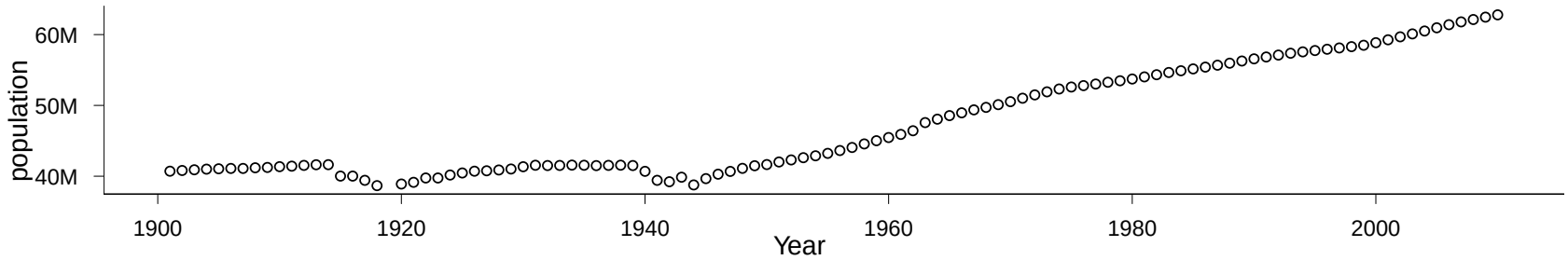
2. Classification *Outskewer*

a) fenêtre temporelle de taille w_o **sur la série temporelle**.
b) chaque valeur est classée w_o fois (chaque fenêtre vote).
c) classe finale : la plus fréquente pour chaque valeur (vote à la majorité simple).

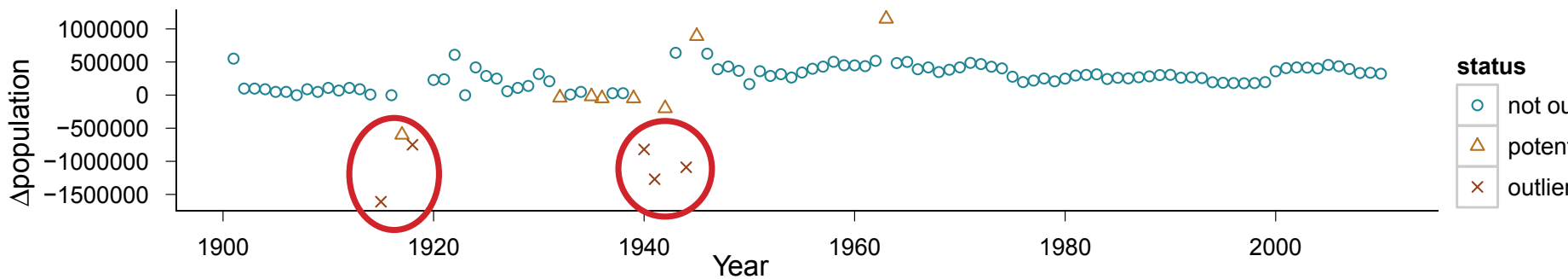
Applicable en temps réel !

Population française au 20^{e} siècle

Nombre d'habitants par an

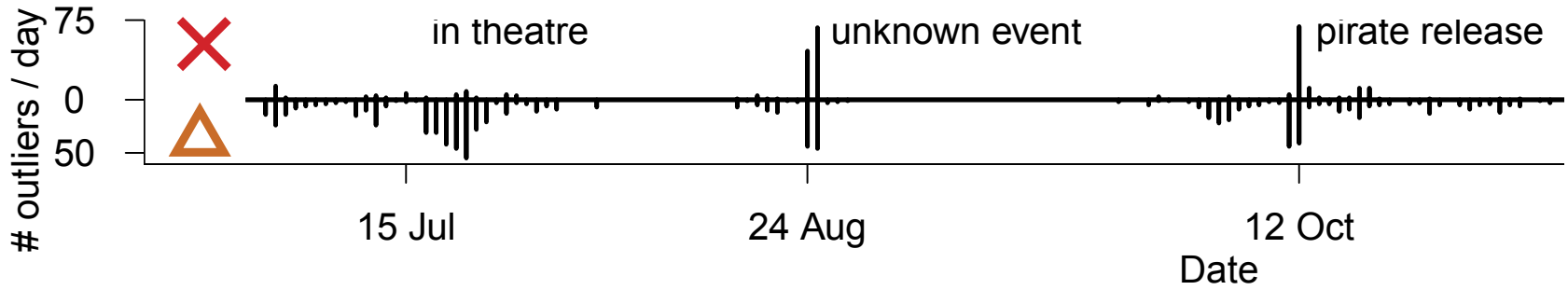


Différence entre les années



Harry Potter sur *eDonkey*

Nombre d'anomalies par jour

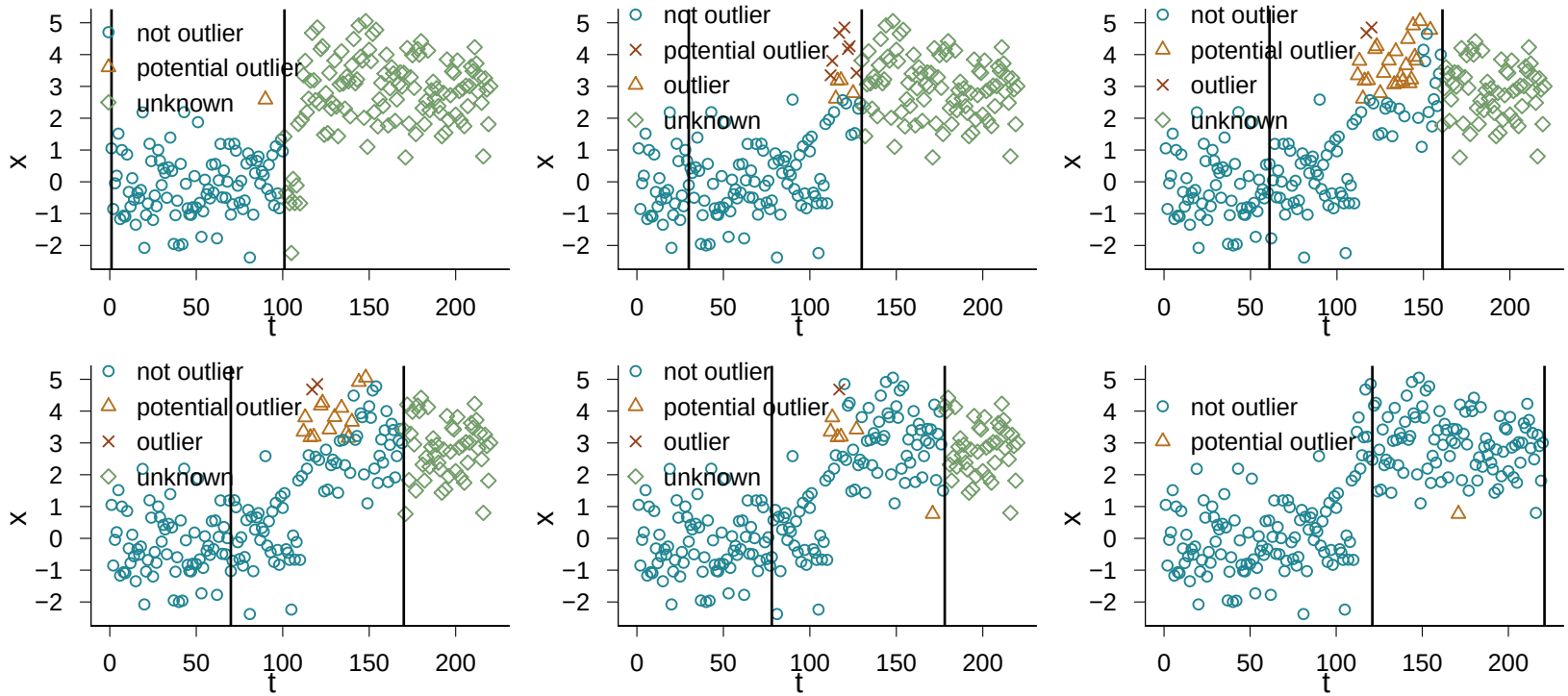


Données :

- recherches sur le réseau P2P *eDonkey*.
- # requêtes contenant "*half blood prince*" par heure, calculées toutes les 10 minutes.
- durant 28 semaines.
- sur 205 millions de requêtes.
- pour 24,4 millions d'adresses IP.
- `http://antipaedo.lip6.fr`

Changements de régime

Video



Données Github.com

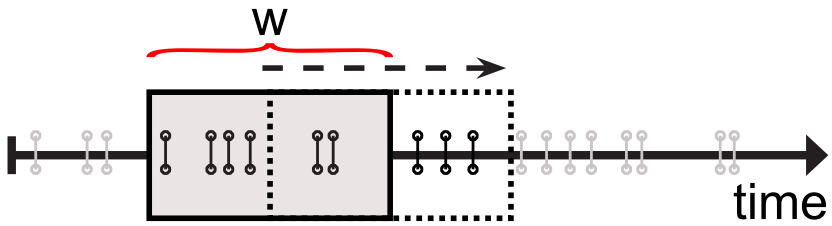
Exemples d'interactions

- <> ajouter du code.
- rapporter un bug.
- ” commenter un bug.
- Aあ éditer le wiki.

"qui contribue à quel projet logiciel"

- 336 000 utilisateurs et projets observés durant 4 mois.
- 2 200 000 interactions enregistrées séquentiellement (avec timestamps).
- "population" complète de Github et non un échantillon.
- `http://www.githubarchive.org`

Quelle unité de temps ?



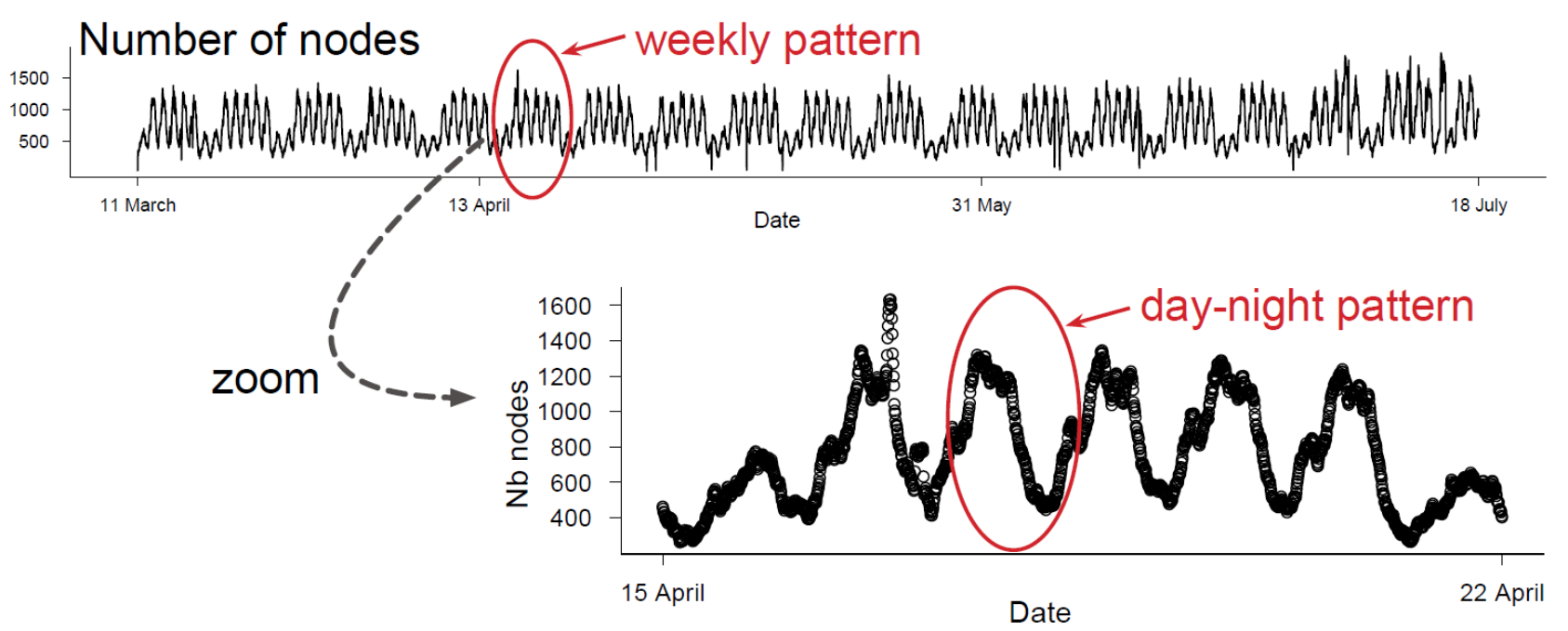
Montres, sablier, horloge atomique : du "temps spatialisé" (E. Klein).

Quelle unité de temps ?

Temps extrinsèque (temps usuel)

Temps mesuré avec des unités comme la seconde, minute, ...

Temps extrinsèque



w =1 heure, i = 5 minutes.

- Cycles journaliers et hebdomadaires.
- Événements difficilement repérables sans modèle.

Quelle unité de temps ?

Temps extrinsèque (temps usuel)

Temps mesuré avec des unités comme la seconde, minute, ...

Révèle habituellement des phénomènes exogènes, ex cycles jour-nuit.

Temps intrinsèque (lié à la dynamique du graphe)

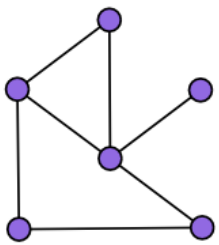
Temps mesuré avec des unités comme la transition de 2 états du graphe. Ex. 1 apparition d'un lien dans un flot de liens.

Meilleur pour révéler des phénomènes endogènes ?

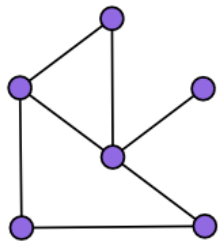
Temps extrinsèque vs temps intrinsèque

Temps **extrinsèque** (absolu)

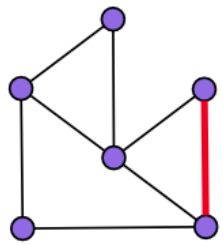
t = 1s



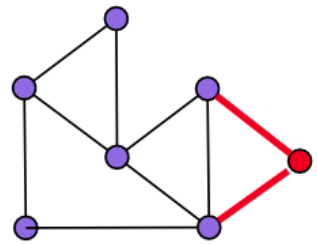
t = 2s



t = 3s

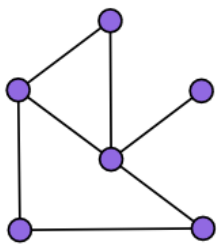


t = 4s



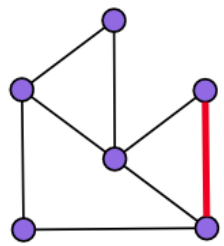
---

t = 1

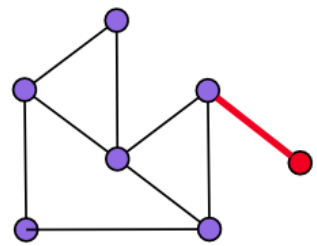


pas de changement dans le graphe

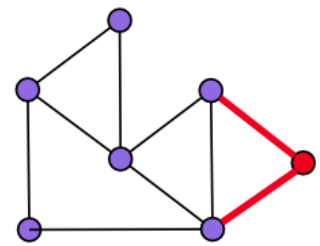
t = 2



t = 3



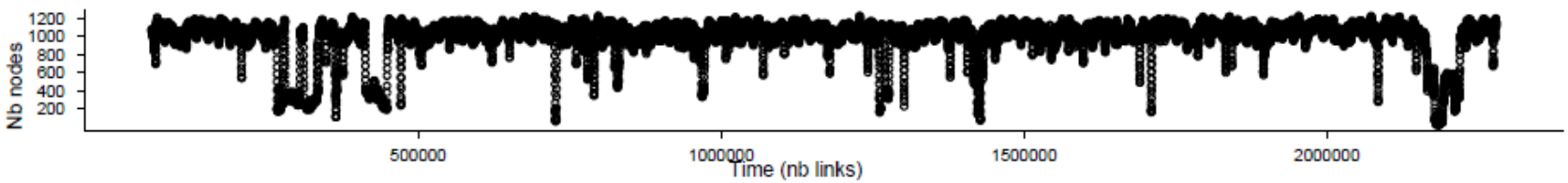
t = 4



Temps **intrinsèque** (lié à la dynamique)

Temps intrinsèque

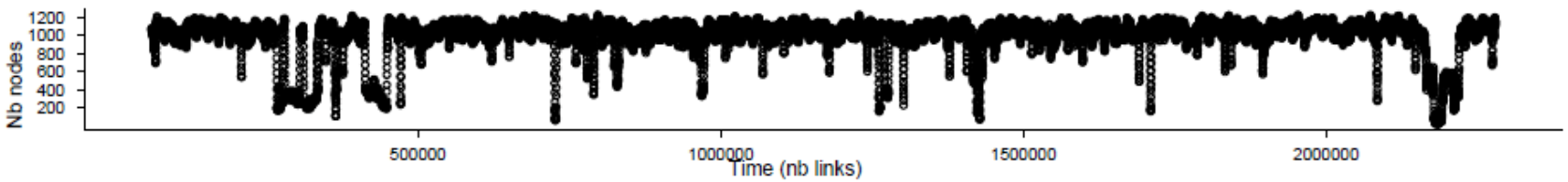
Haute résolution



Disparition de la dynamique exogène.

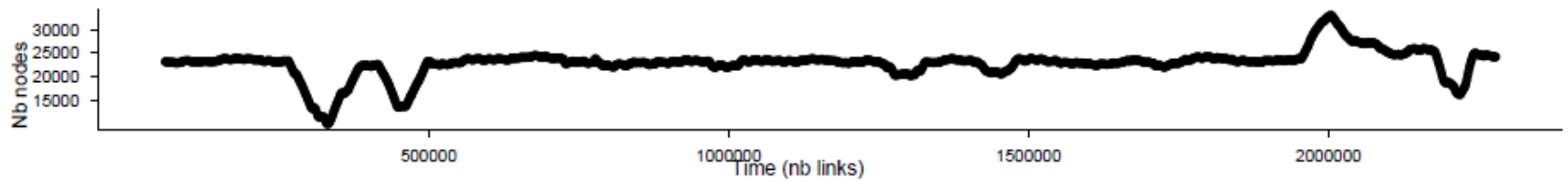
Taille de fenêtre ?

Haute résolution



$w = 1000$ liens, $i = 100$ liens.

Basse résolution



$w = 50,000$ liens, $i = 1000$ liens.

Conclusions sur la détection automatique d'événements

Méthode

- Applicable en temps réel

Unité de temps

- Notion d'unité de temps : invention humaine
- Temps intrinsèque : lié à la dynamique du graphe
- Travaux en cours

Taille de fenêtre

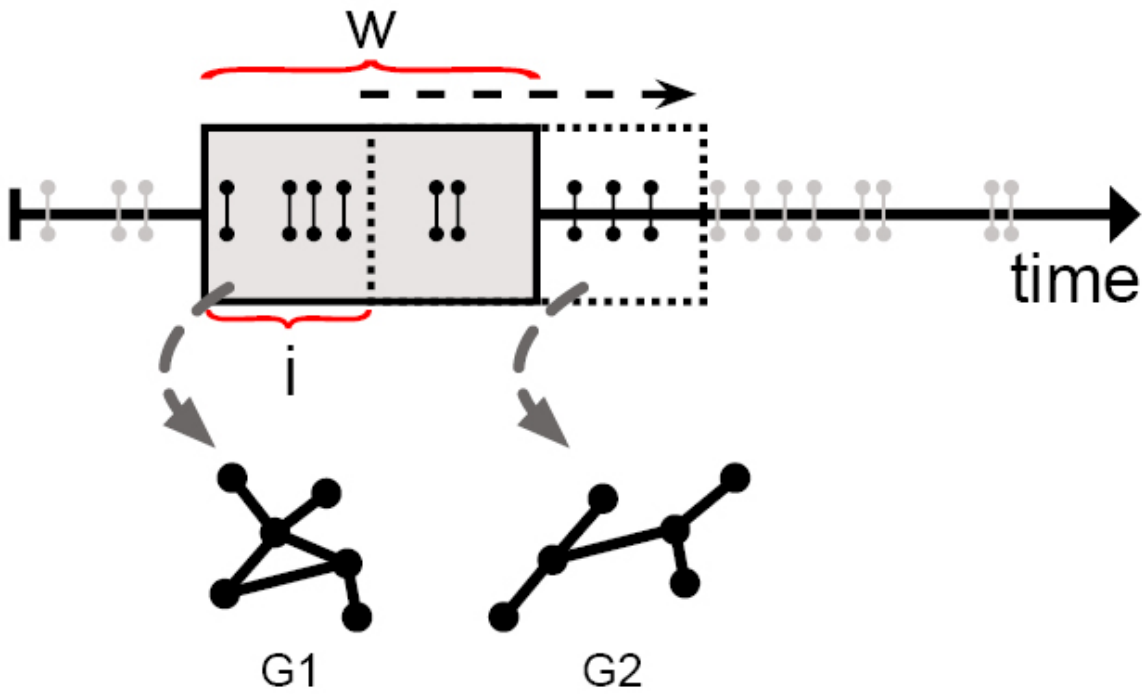
- Semble liée à la résolution des événements
- Pas de résolution optimale ?

Publications : [INFORSID 13] [RCIS 13] [WebSci 13] [ASONAM 13]

Contributions

1. Détection automatique d'anomalies.
   - nouvelle méthode statistique : *Outskewer*
   - validation expérimentale
2. Détection automatique d'événements.
   - notion de fenêtre temporelle
   - unité de temps : intrinsèque vs extrinsèque
   - largeur de fenêtre
3. **Cadre exploratoire et étude de cas.**
   - **méthodologie unifiée : statistiques et visualisation**
   - **prototype**
   - **application à Github.com**

Méthodologie unifiée



1. Définir une fenêtre temporelle de taille w liens.
2. Extraire le graphe de chaque fenêtre à intervalle i.
3. Calculer une propriété sur chaque graphe.
4. Détecter des événements dans la série temporelle résultante.
5. Analyser ces événements par la visualisation de graphes.

Analyse visuelle des événements

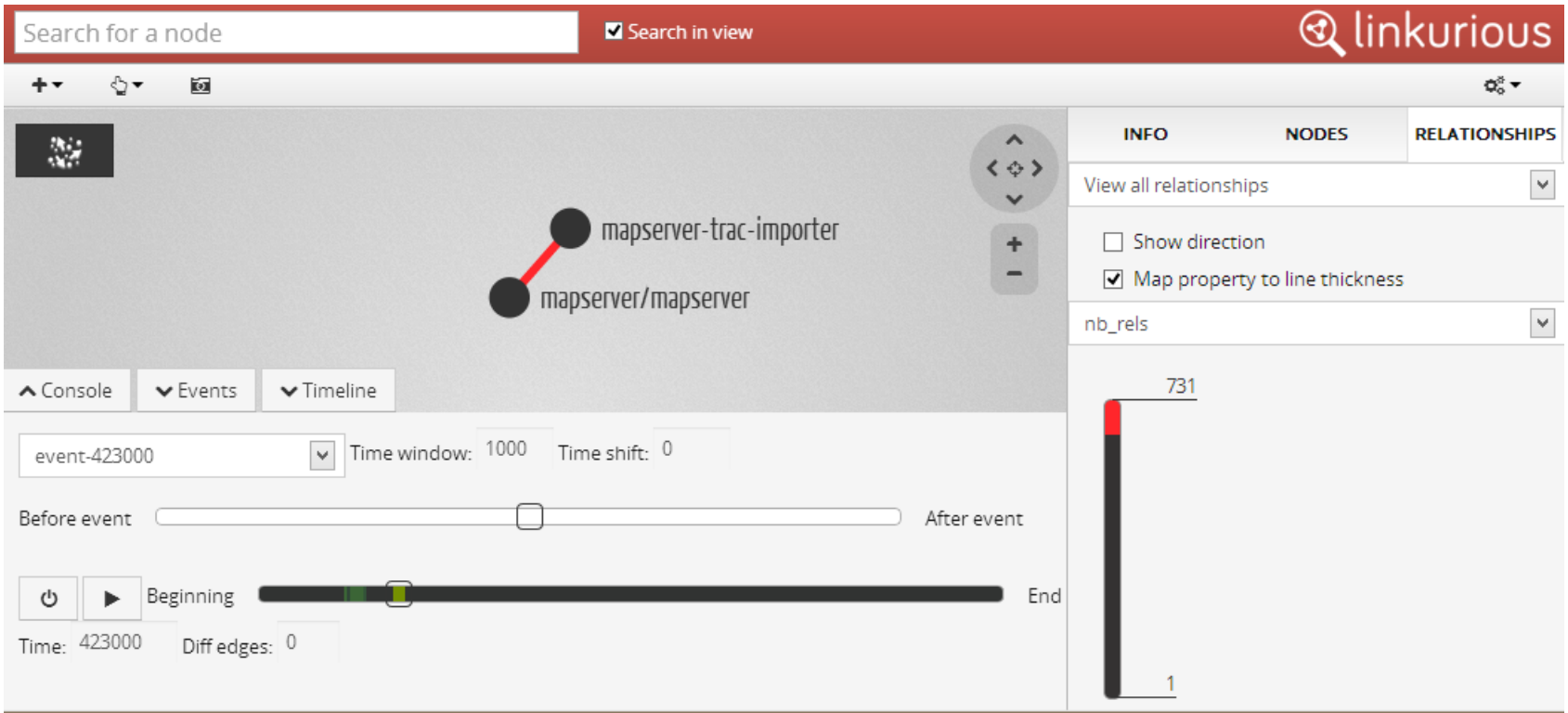
1. Sélectionner un événement détecté par *Outskewer*.
2. Identifier un motif anormal dans le sous-graphe correspondant de taille $w_v \leq w$.
3. Interpréter l'événement. Vérifier que le motif est unique au moyen d'un *diagramme d'activité*.
   - Si motif unique : événement validé
   - Sinon événement non validé, chercher un autre motif.

Diagramme d'activité

- S'applique à un ensemble de nœuds.
- Fréquence des liens entre cet ensemble et le graphe.
- Axe horizontal : temps.
- Couleur : intensité de l'activité (noir ou nuances de vert).

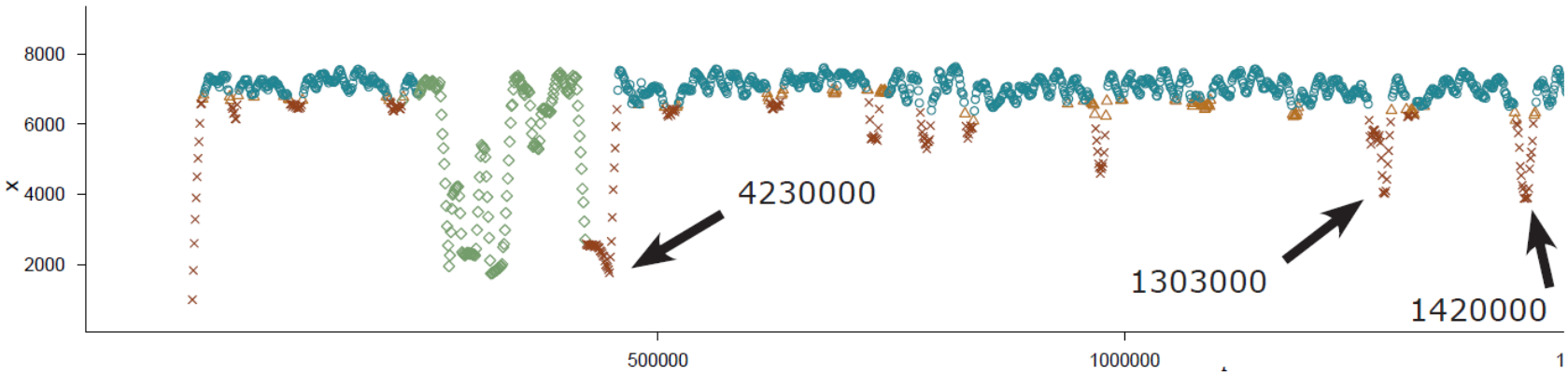
Diagramme d'activité du projet "mxcl/homebrew" (github.com). Ce projet reçoit des contributions tout au long de la période de capture des données.

Prototype



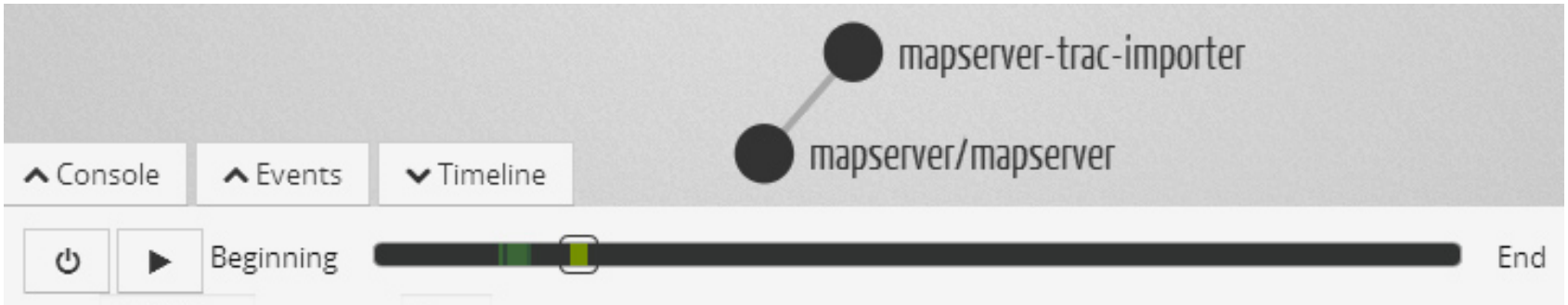
Basé sur le logiciel Linkurious

Étude de cas : Github.com



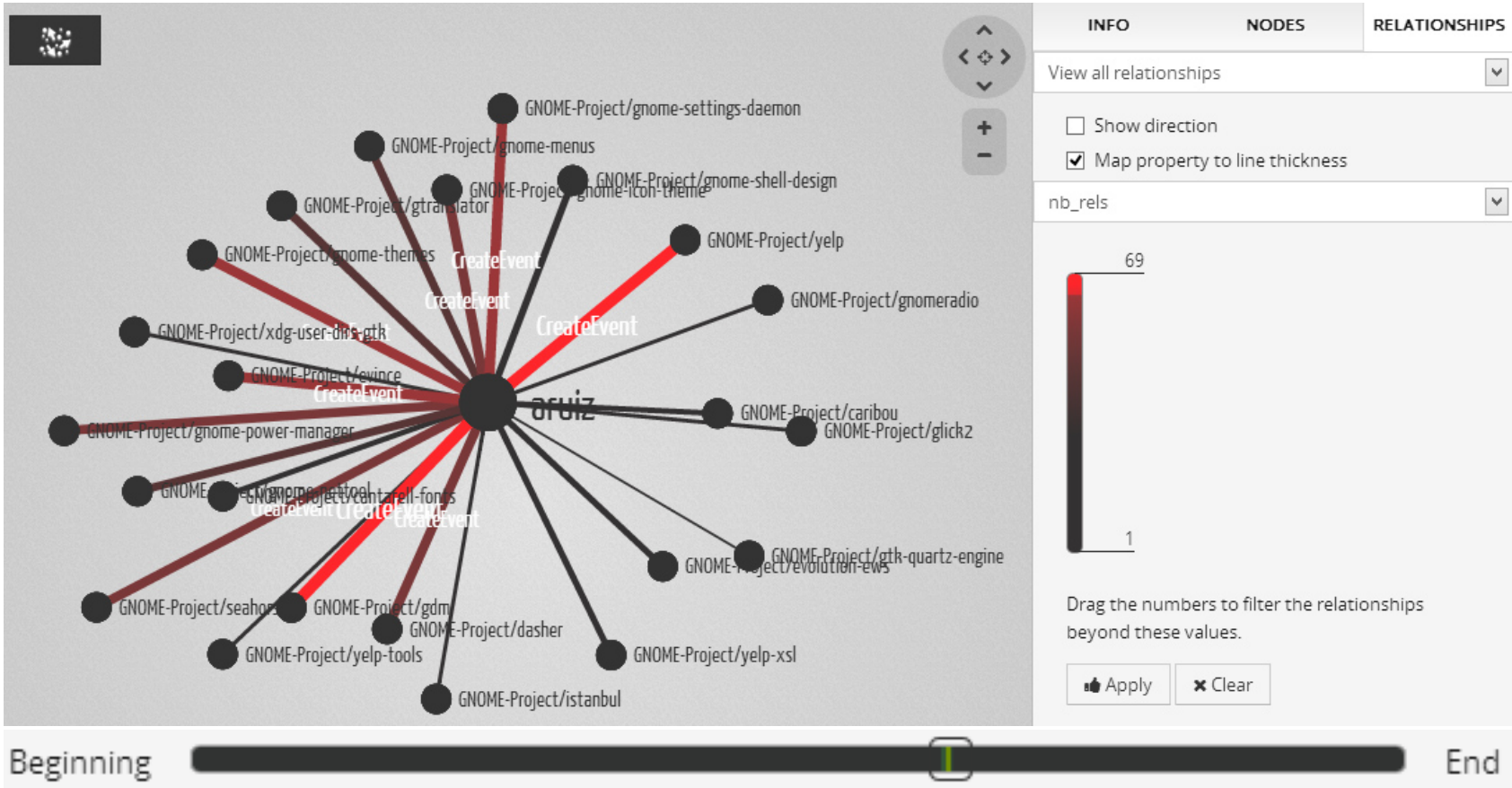
Série temporelle du nombre de noeuds uniques sur le flot de liens de Github, avec $w = 10000$ liens et $w_o = 200$ liens.

Exemple d'événement validé



- Lien très récurrent entre l'utilisateur "mapserver-trac-import" et le projet "mapserver/mapserver" lors de l'événement 423000 (période en vert clair).
- Migration d'un dépôt logiciel de Trac vers Github grâce à un utilisateur "bot".
- A investiguer : autre période d'activité plus tôt (vert peu intense).

Exemple d'événement validé



Visualisation des liens de l'utilisateur "aruiz" lors de l'événement 1420000 avec $w_v = 1000$ et son diagramme d'activité.

Exemples d'événements rejetés

Diagramme d'activité entre l'utilisateur "pjcozzi" et le projet "AnalyticalGraphicsInc/cesium" lors de l'événement 1303000. Le motif est récurrent.

Publications : [CHItaly 13] [EGC 14] [chap. Gephi Comp.Net 14]

Conclusions générales

Détection automatique d'anomalies

- Méthode statistique automatique : *Outskewer*.
- Validation sur des données réelles et simulées.

Détection d'événements dans des grands flots de liens

- Paramètres de la fenêtre temporelle et impact sur les événements.
- Analyse des événements par fouille visuelle interactive.
- Validation possible et interprétation des événements.

Cadre exploratoire

- Prototype interactif.
- Étude de cas sur Github.com

Perspectives

- Complexité algorithmique et optimisation.
- Comparer à d'autres méthodes. Hypothèses de départ ?

- Approfondir l'étude du temps intrinsèque
- Unité représentative de la dynamique ? ex : triangle observé

- Évaluer par des utilisateurs réels.
- Généraliser à d'autres jeux de données
  ex : télécom, détection de fraude.

KNOWeSCAPE

EU COST Action 2013-2017

Start-up francilienne

Publications

ASONAM 12 Heymann, Latapy and Magnien. Outskewer : Using Skewness to Spot Outliers in Samples and Time Series.

CHItaly 13 Uboldi et al., Knot : an Interface for the Study of Social Networks in the Humanities.

INFORSID 13 Heymann and Le Grand. Suivi de la Dynamique Intrinsèque des Interactions entre Utilisateur et SI.

RCIS 13 Heymann and Le Grand. Monitoring User-System Interactions through Graph-Based Intrinsic Dynamics Analysis.

WebSci 13 Heymann and Le Grand. Towards A Redefinition of Time in Information Networks ?

ASONAM 13 Albano et al., A Matter of Time - Intrinsic or Extrinsic - for Diffusion in Evolving Complex Networks.

chap. Gephi Comp.Net 14 Heymann and Le Grand. Exploratory Network Analysis : Visualization and Interaction. à paraître dans Complex Networks and their Applications.

Encycl. SNAM 14 Heymann, Gephi. à paraître dans Encyclopedia of Social Networks and Mining.

EGC 14 Heymann and Le Grand. Investigation visuelle d'événements dans un grand flot de liens. à paraître à EGC 2014.