Monitoring User-System Interactions through Graph-Based Intrinsic Dynamics Analysis

Sébastien Heymann, Bénédicte Le Grand

Emails: Sebastien.Heymann@lip6.fr, Benedicte.Le-Grand@univ-paris1.fr

Monitoring user-system interactions

What type of user-system interactions?

- user-invoked services in information systems
- social networks
- ...

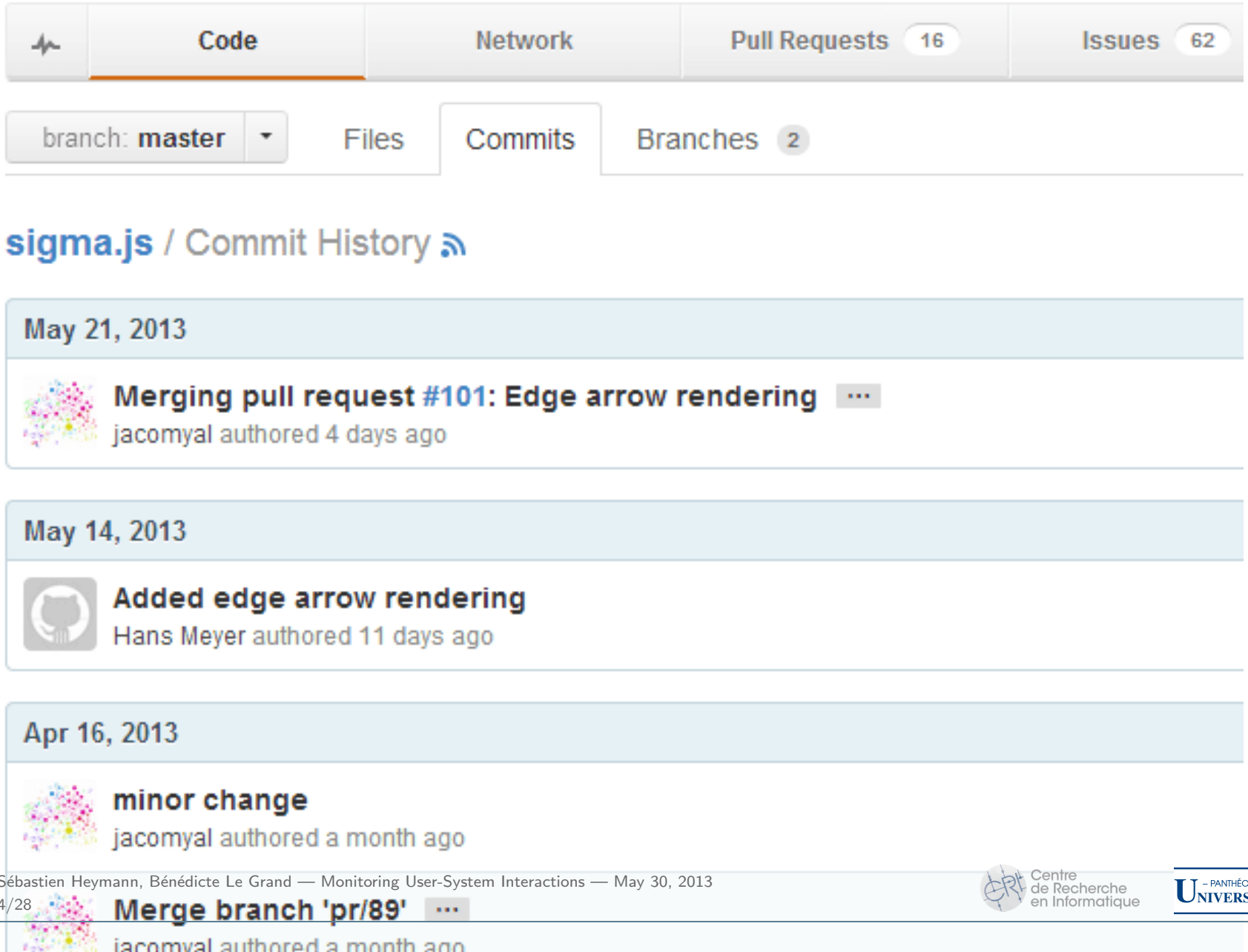
What kind of monitoring?

- **discovery**
- conformance
- model improvement

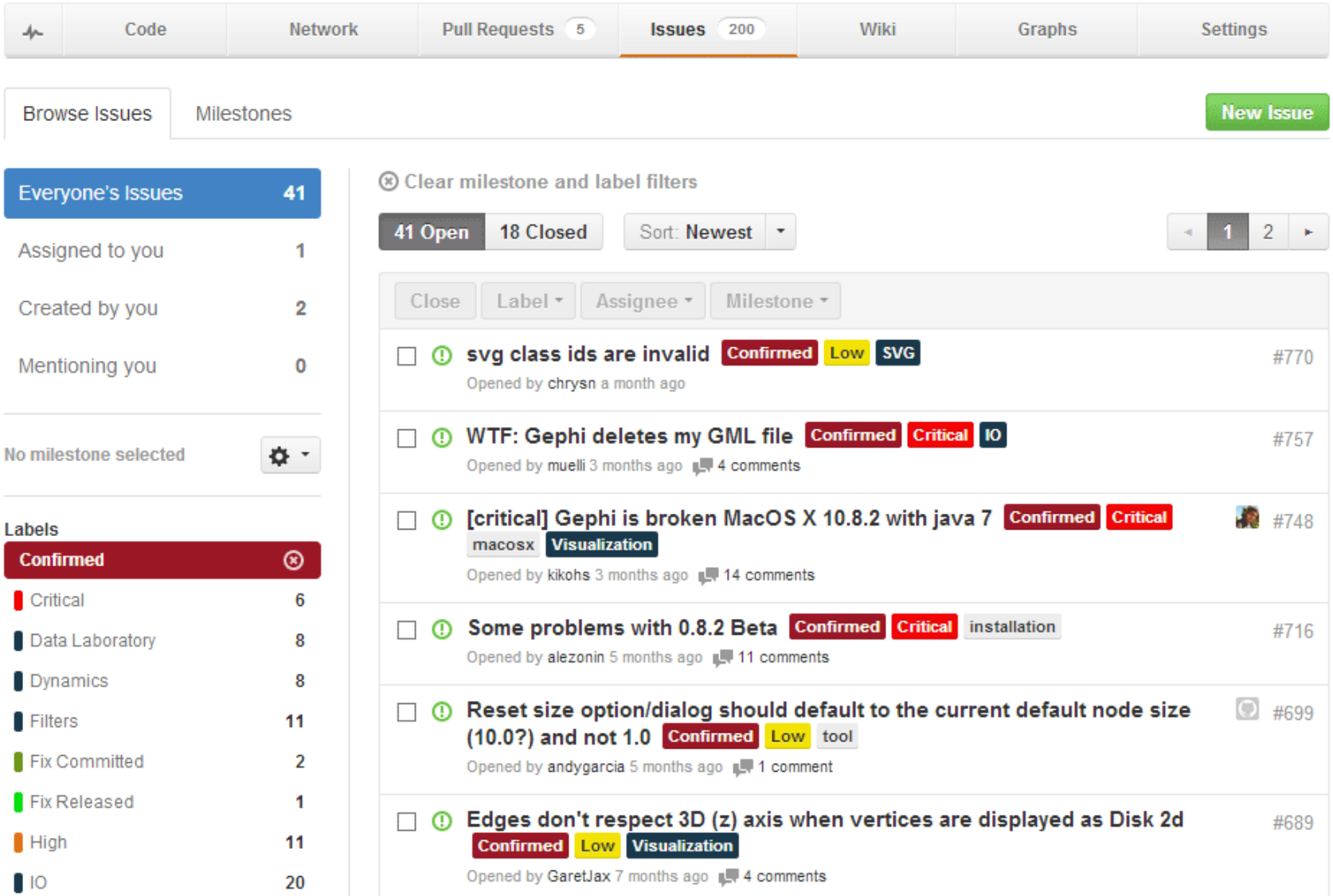
Our ultimate goal: automatic and real-time anomaly detection.

Studied social network

Github interaction: code commit



Github interaction: bug report



Collected Dataset

Interactions examples

<> commit code / merge repositories.

open / close bug reports.

” comment on bug reports.

Aあ edit the repository wiki.

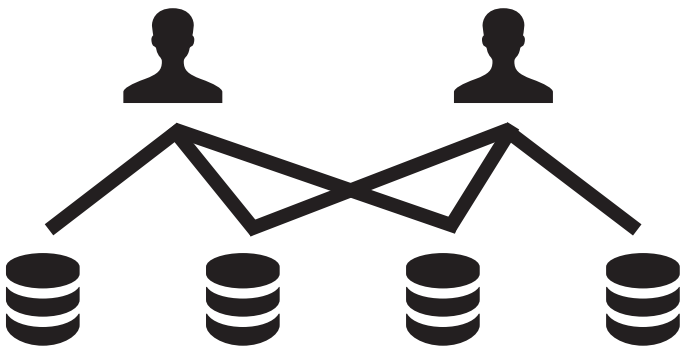
”who contributes to which source code repository”

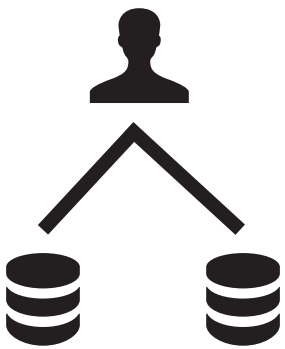
- 336 000 users and repositories monitored during 4 months.
- 2.2 million interactions recorded sequentially with timestamps.

Log trace sample

User, user, repository, event, timestamp
lukearmstrong, fuel, core, IssuesEvent, 1341420003
Try-Git, clarkeash, try_git, CreateEvent, 1341420006
uGoMobi, jquery, jquery-mobile, IssuesEvent, 1341420009
jexp, neo4j, java-rest-binding, IssueCommentEvent, 1341420011
HosipLan, nette, nette, PullRequestEvent, 1341420152

Bipartite graph



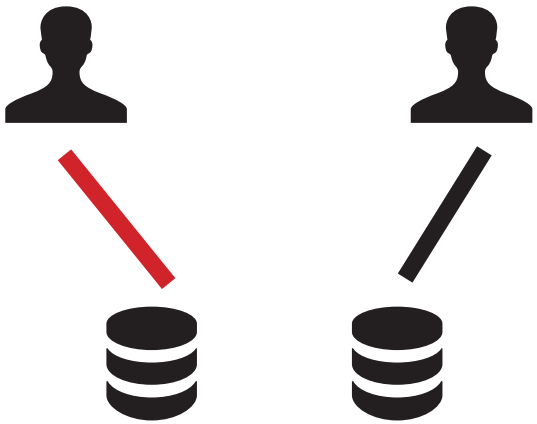


$\top$: users

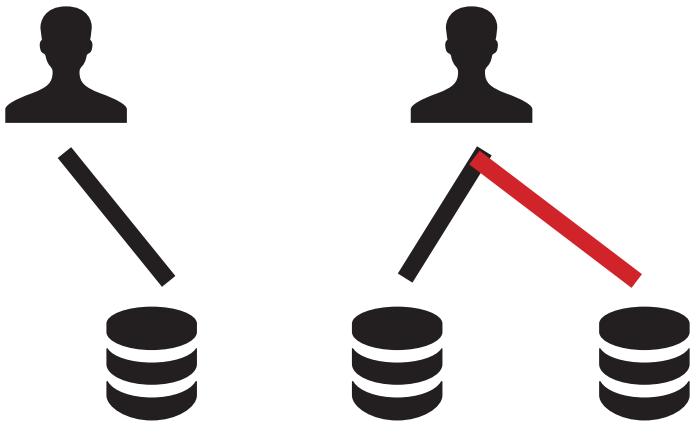
$\bot$: repositories

Links appear over time

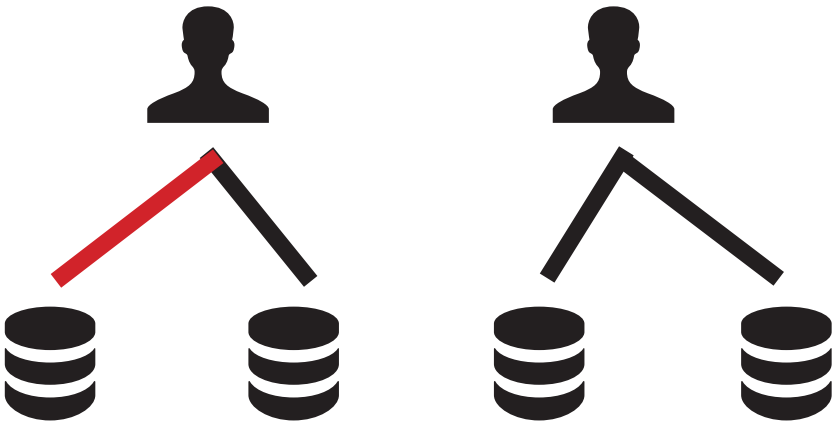
Links appear over time



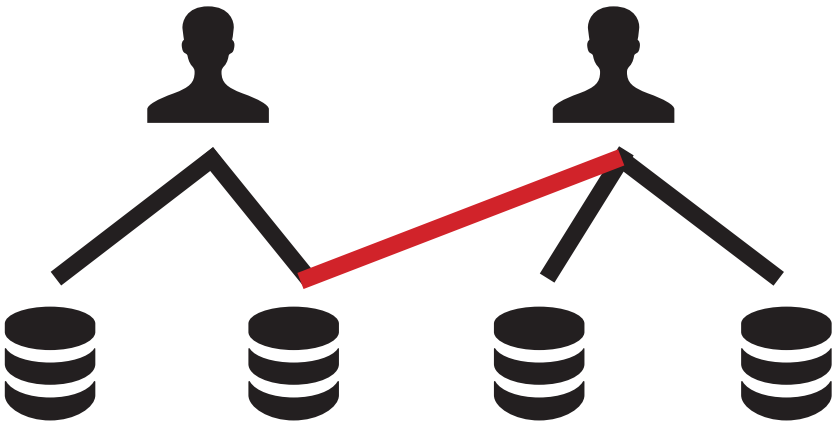
Links appear over time



Links appear over time

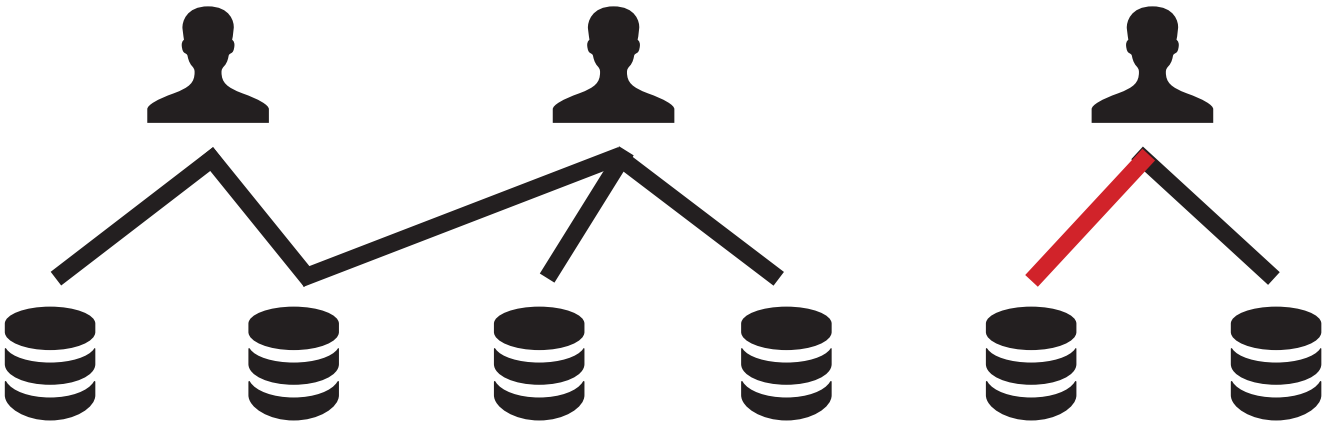


Links appear over time

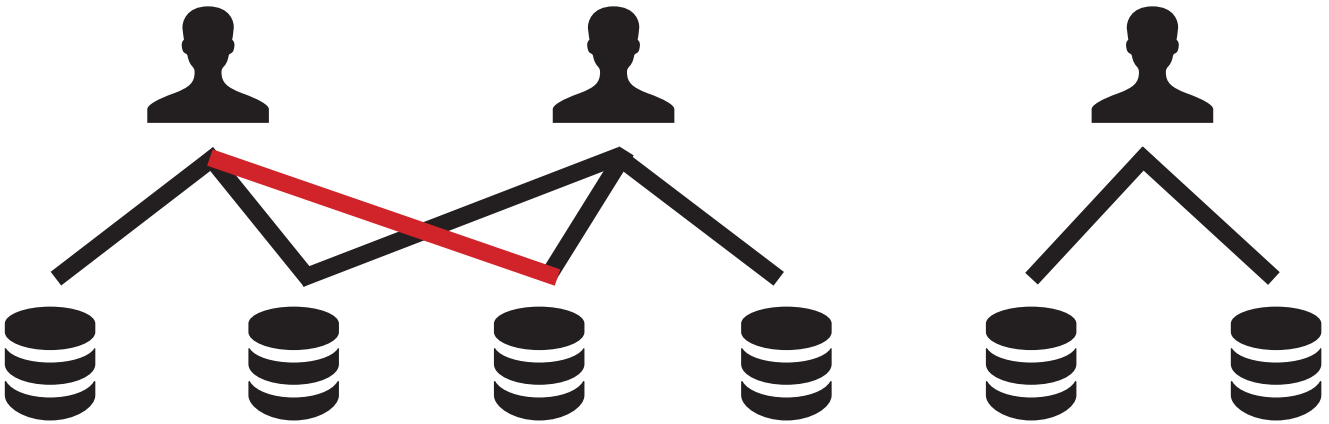


Links appear over time

Links appear over time



Links appear over time



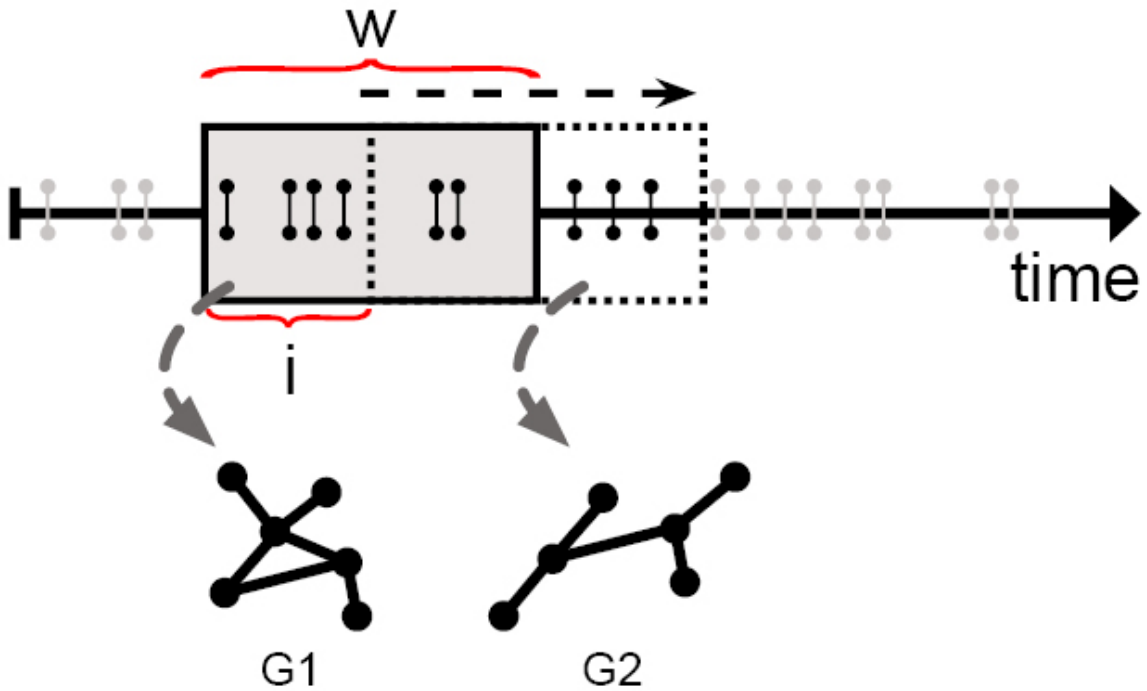
Detection of statistically abnormal links dynamics?

Model of links dynamics?

Link prediction?

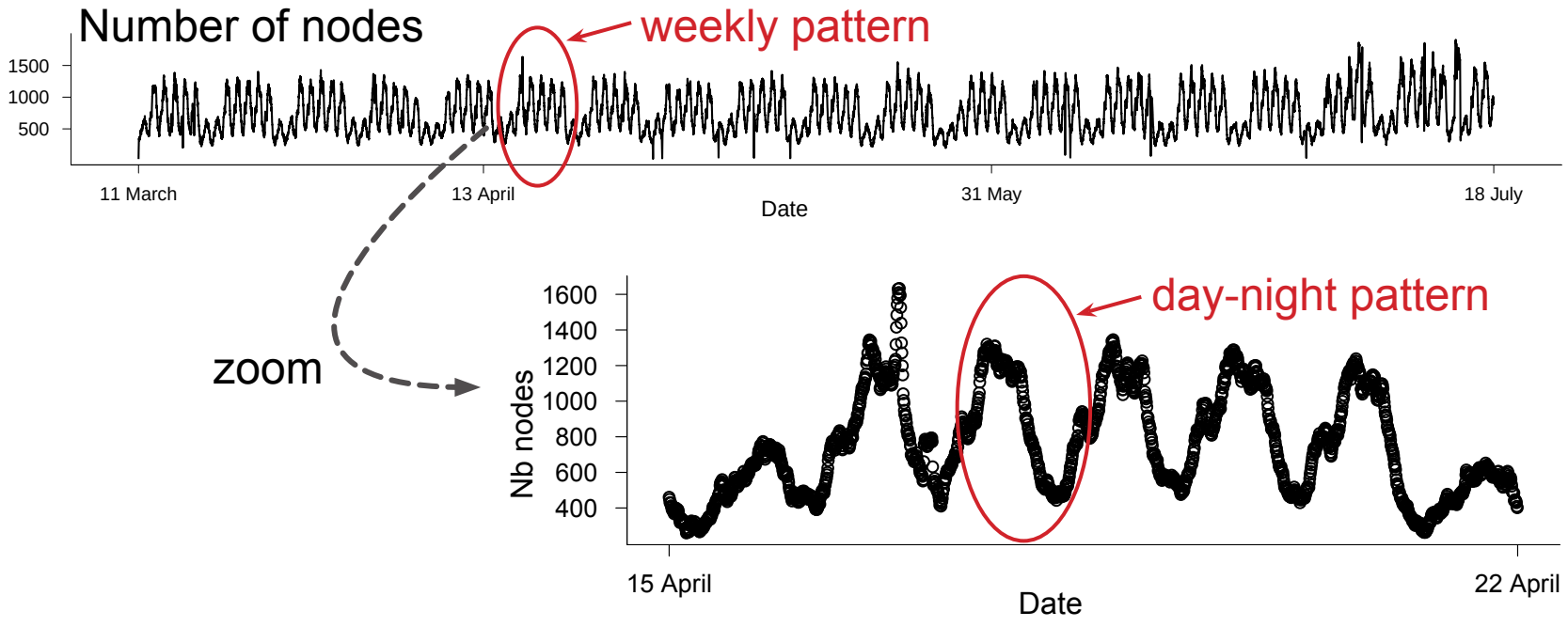
Methodology

1. Order links by timestamp.
2. Define a sliding window of width w (time unit?).



3. Extract the bipartite graph from each window at interval i.
4. Compute an appropriate property on each graph.
5. Analyze the time series.

Example



w =1 hour, i = 5 minutes.

Question: don't temporal patterns hide information?

Notions of time

Extrinsic time (real time)

Time measured in units such as seconds.

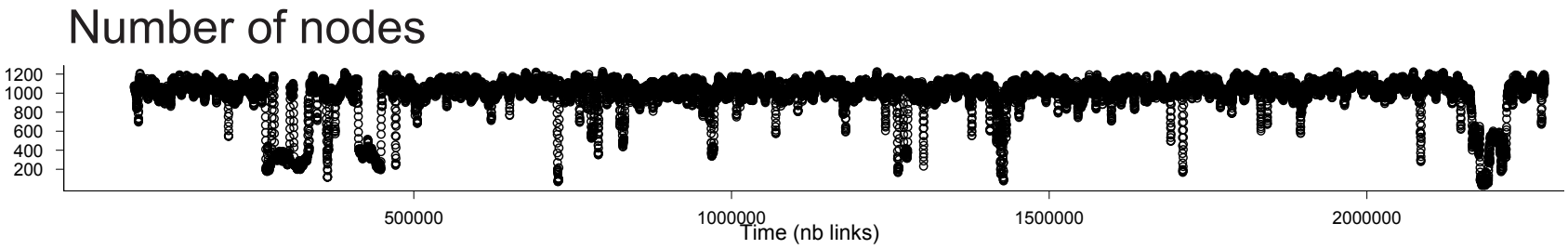
Good at revealing exogenous phenomena, e.g. day-night patterns.

Intrinsic time (related to graph dynamics)

Time measured in units such as the transition of two states in the graph.

Better at revealing endogenous phenomena independently from the graph dynamics?

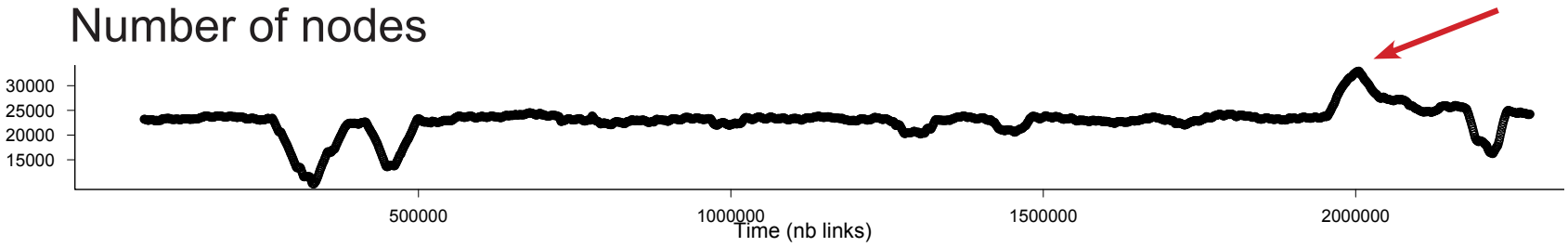
Window width: high resolution



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

:) Additional observation

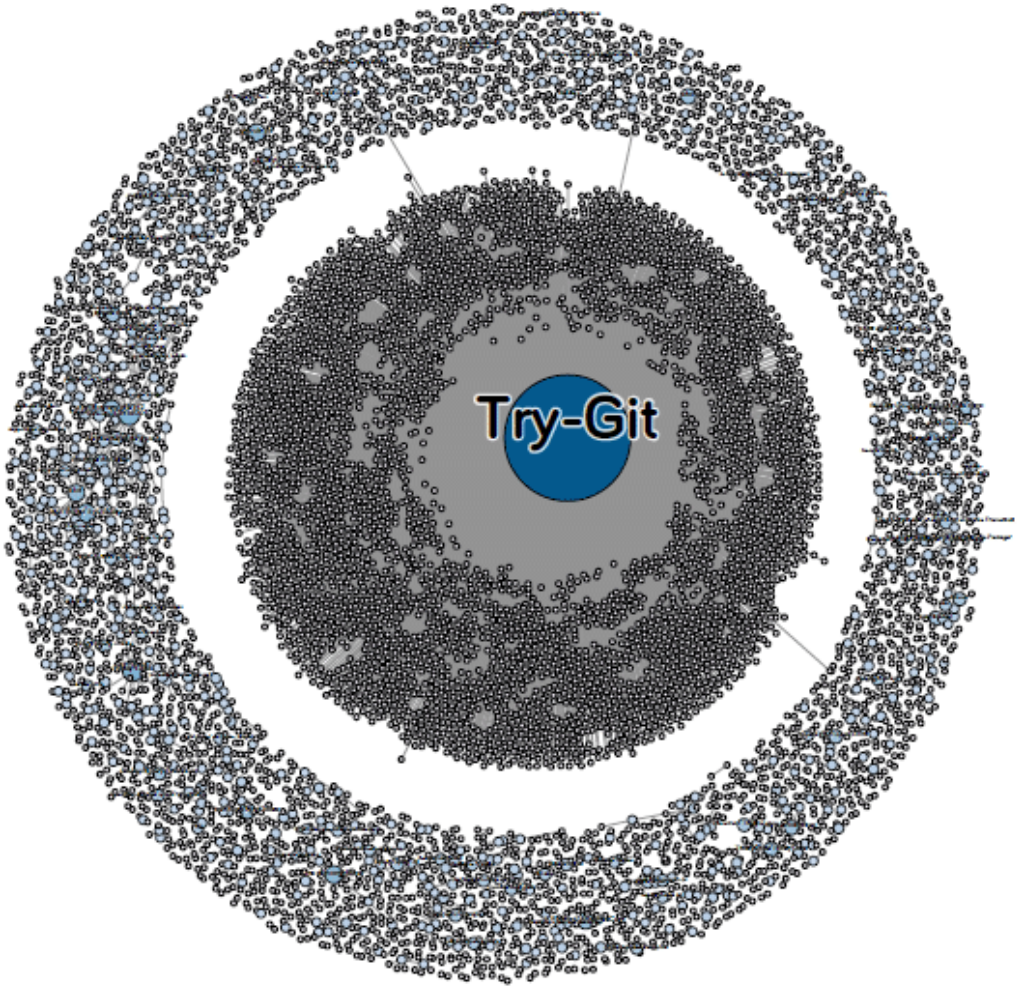
Window width: lower resolution



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

:) No need for high resolution

Event validation



In the sub-graph of 8,370 nodes and 10,000 links at the time of the event, one node has a high number of links:

Try-Git interacts with 4,127 users (over 5,000).

Visualization of the sub-graph: connected nodes are closer, disconnected nodes are more distant.

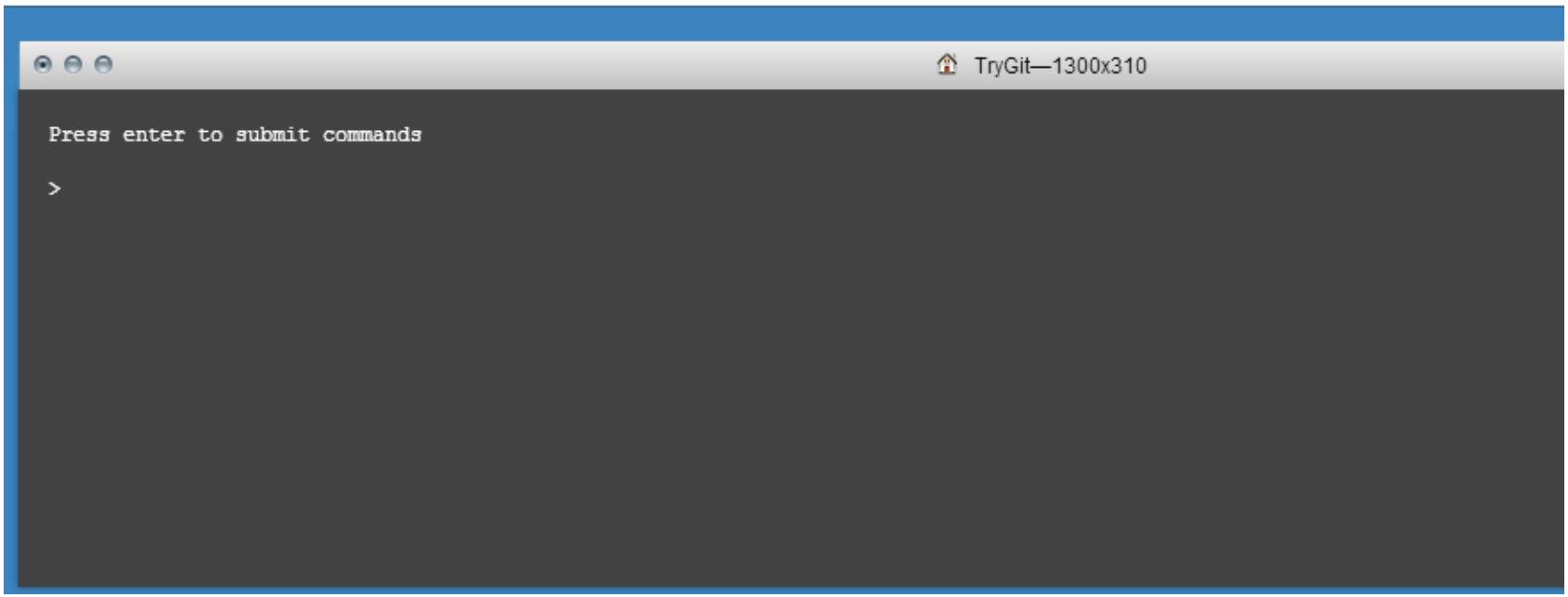
http://try.github.io

1.1 · Got 15 minutes and want to learn Git?

Git allows groups of people to work on the same documents (often code) at the same time, and without stepping on each other's toes. It's a distributed version control system.

Our terminal prompt below is currently in an **octobox** directory. To initialize a Git repository here, type the following command:

```
git init
```



Towards automatic anomaly detection

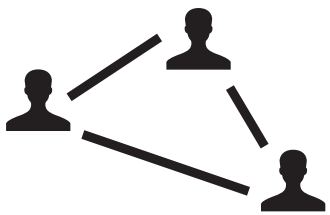
Need for more elaborate properties, like:

Internal links

Their removal does not change the projection of the graph for a given set of nodes, either $\top$ or $\bot$.

G

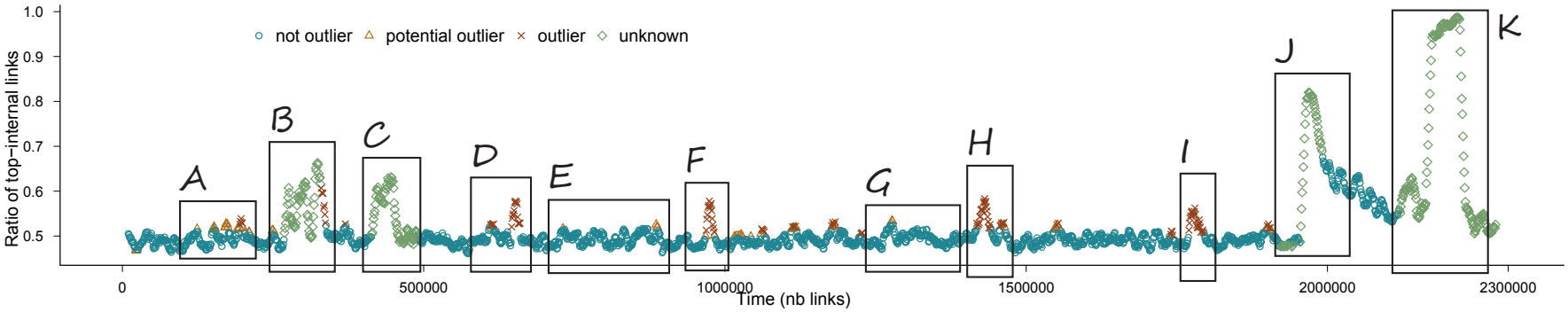
G' = G - (red link)



$G'_T = G_T$

Results

Ratio of ⊤-internal links



$w = 10,000$ links, $i = 1000$ links.

Color = outlier class using the automatic Outskewer method*.

** S. Heymann, M.Latapy and C. Magnien. Outskewer: Using Skewness to Spot Outliers in Samples and Time Series, IEEE ASONAM 2012*

Conclusion

Contributions

- Graph-based methodology to monitor user-system interactions
- Intrinsic time unit avoids exogeneous patterns impact
- Smaller windows not necessarily optimal
- Checked relevance of detected events

Applicable in other contexts

- Client-server architectures
- Processes-messages graphs
- File-provider graphs
- User-invoked services in information systems

Future work

- Which property for anomaly detection?
- Models of interaction dynamics
- Link prediction

Questions?

Monitoring User-System Interactions through
Graph-Based Intrinsic Dynamics Analysis

<sebastien.heymann@lip6.fr>

Thank You!

Monitoring User-System Interactions through
Graph-Based Intrinsic Dynamics Analysis

<sebastien.heymann@lip6.fr>

Backup Slides

Statistically significant anomalies

General definition

Values which deviate remarkably from the remainder of values (Grubbs, 1969)

Outskewer method*:

Our definition

Extremal value which skews a distribution of values.

** Heymann, Latapy and Magnien. Outskewer: Using Skewness to Spot Outliers in Samples and Time Series, IEEE ASONAM 2012*

Skewness coefficient

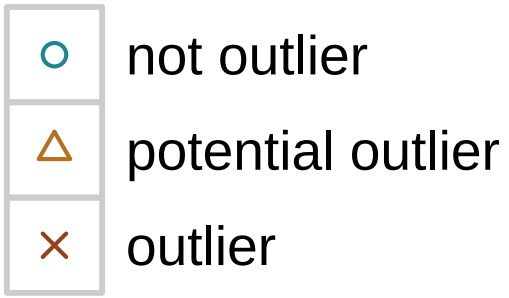
$$\gamma = \frac{n}{(n-1)(n-2)} \sum_{x \in X} \left(\frac{x - \text{mean}}{\text{standard deviation}} \right)^3$$

Example of skewed distributions.

It is sensitive to extremal values (min/max) far from the mean !

Automatic anomaly detection

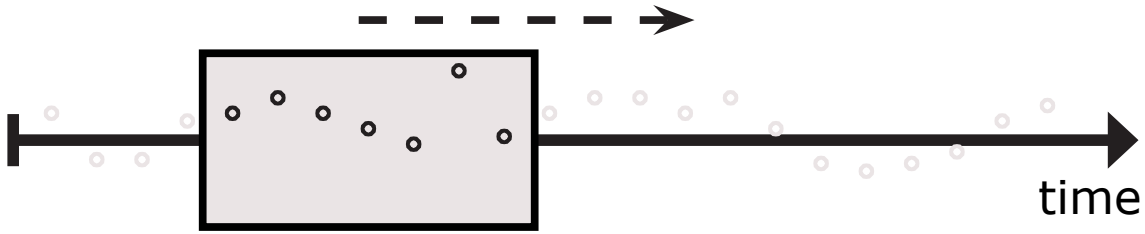
Outskewer classifies each value as:



or 'unknown' for heterogeneous distributions of values.

Event detection in time series

On a **sliding window** of size w, each value of X is classified w times.
The final class of a value is the one that appears the most.



Why Outskewer?

- claims no strong hypothesis on data
- 1 parameter: the time window width
- ignores regime changes (shifts in normality)
- can be implemented on-line.