Stream graphs and link streams for the modelling of interactions over time, and application to contact analysis

Tiphaine Viard

tiphaine.viard@riken.jp

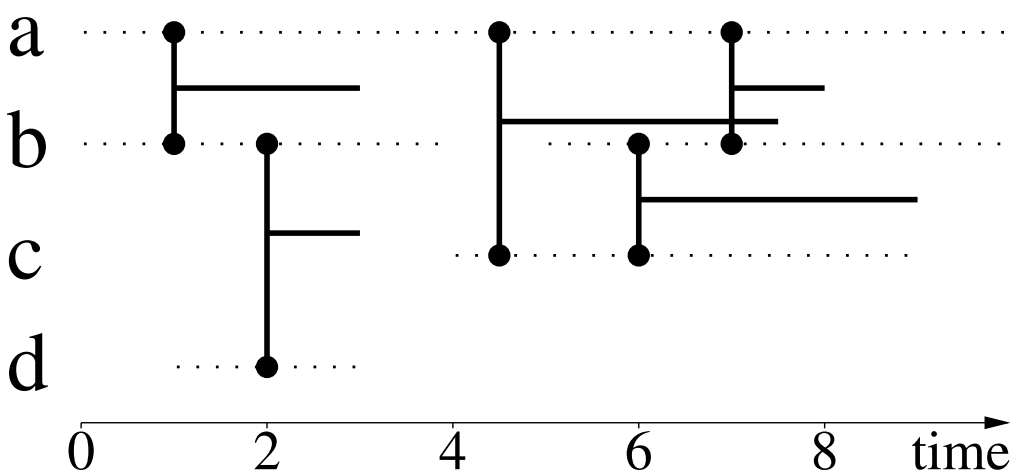
September 20th, 2018

Our topic: interactions over time

numerous examples

email exchanges, IP traffic, online transactions,
face-to-face contacts, phone calls...

interactions over time

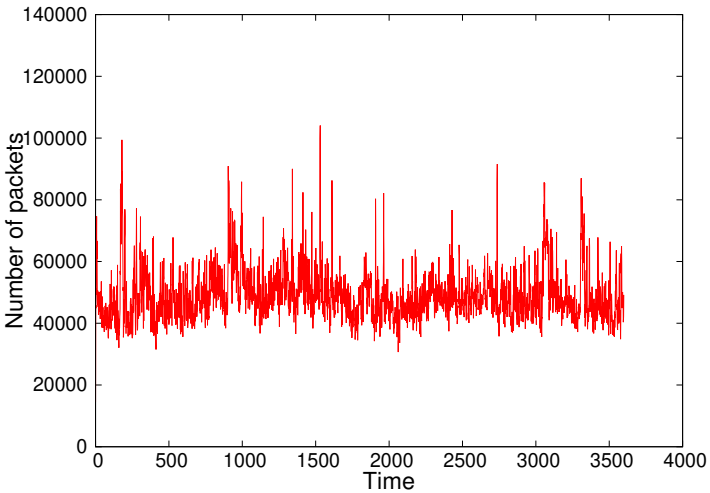


$\ell = (t, uv)$

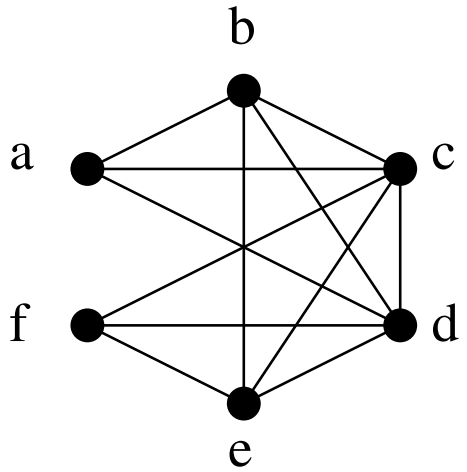
$t \in [\alpha, \omega]$: time
$u, v \in V$: nodes

Context

Time (signal processing) :



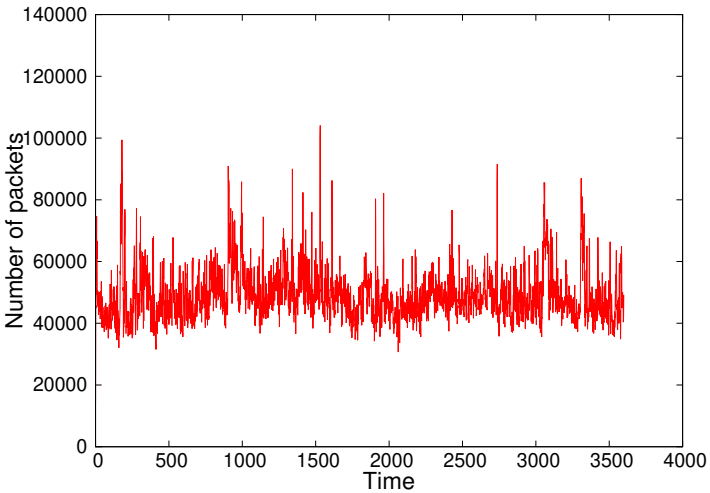
Structure (graph theory):



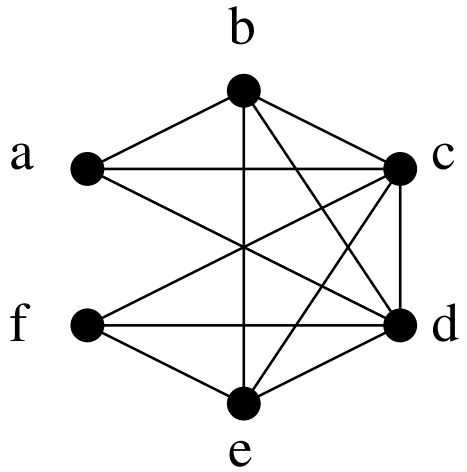
Information loss

Context

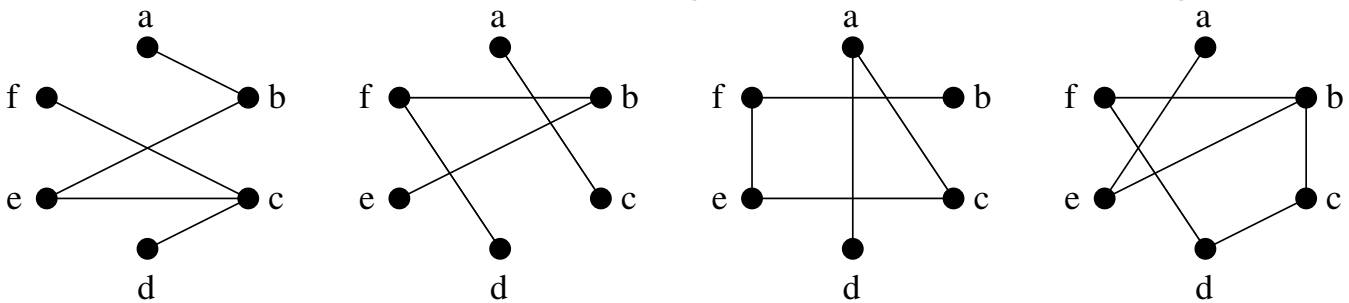
Time (signal processing) :



Structure (graph theory):



"Time" + structure (sequence of graphs)



T=[0,5) T=[5,10) T=[10,15) T=[15,20)

Information loss

Goal

Assessment : progress **limited** by **fundamental** locks

$\rightarrow$ Loss of information, inadequate formalism...

Our goal :
A language for interactions
comparable to graph theory
for networks

- **Simple** and **intuitive**
- Generalizes **graphs** and **signal**
  - degree? clustering? autocorrelation? Fourier transform? ...?
- Allows **applicative progress**

Outline

Formalism

- **Stream graphs**
- **Clusters**
- **Density**
  - **Neighborhood** and degrees
  - **Cliques**
  - **Clustering coefficient**, transitivity
  - Quotient stream
  - k-cores
- Paths
  - Accessibility
  - Connectedness
  - Centralities
  - Trees and cascades

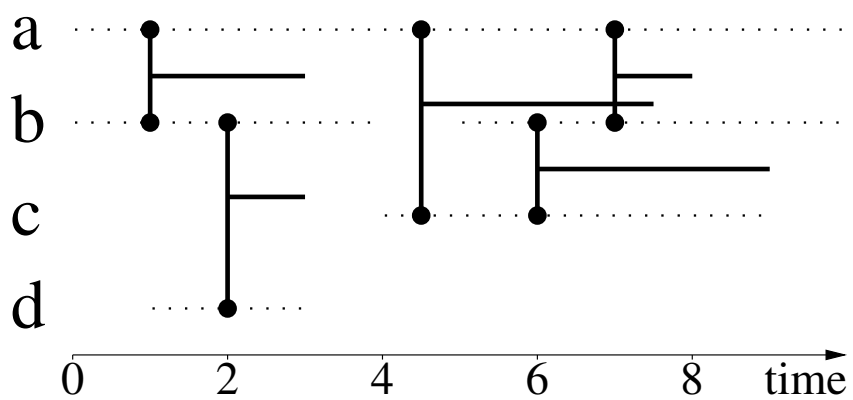
Data analysis

- Face-to-face contacts
  - cliques
- Emails
  - threads = communities?
- IP traffic
  - Bipartite stream graphs
  - Cliques for anomaly detection
- MovieLens

Stream graphs

$S = (T, V, W, E)$

$T = [\alpha, \omega]$ $V = \{u\}$ $W = \{(t, v)\}$ $E = \{(t, uv)\}$



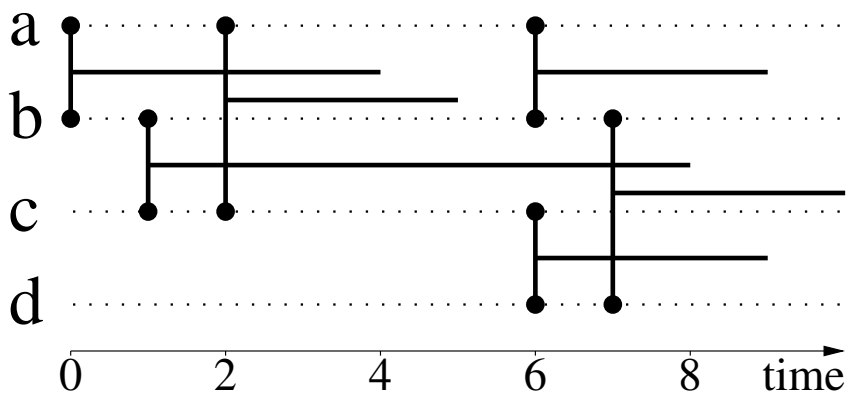
c is present from time 4 to time 9: $\{c\} \times [4, 9] \in W$
a interacts with b from time 1 to time 3: $\{ab\} \times [1, 3] \in E$
...

T_u = presence of node u
$\rightarrow T_b = [0, 4] \cup [4.5, 10]$
T_{uv} = presence of link uv
$\rightarrow T_{ab} = [1, 3] \cup [7, 8]$

Stream graphs

$S = (T, V, W, E)$

$T = [\alpha, \omega]$ $\quad V = \{u\}$ $\quad W = \{(t, v)\}$ $\quad E = \{(t, uv)\}$



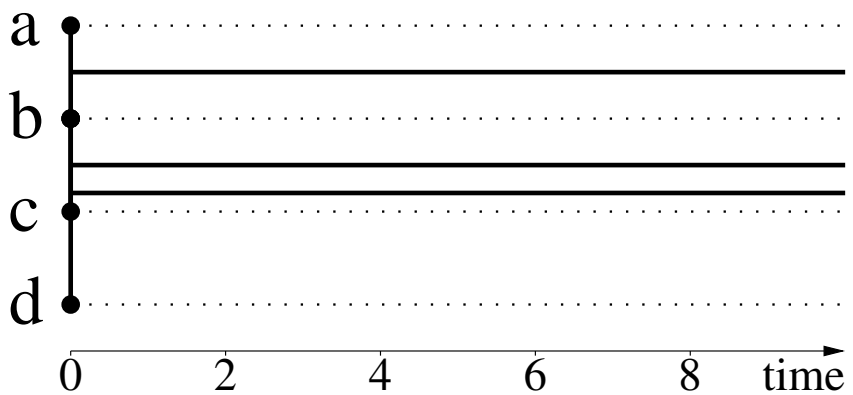
c is present from time 4 to time 9: $\{c\} \times [4, 9] \in W$
a interacts with b from time 1 to time 3: $\{ab\} \times [1, 3] \in E$
...

T_u = presence of node u
$\rightarrow T_b = [0, 4] \cup [4.5, 10]$
T_{uv} = presence of link uv
$\rightarrow T_{ab} = [1, 3] \cup [7, 8]$
If $\forall v, T_v = T$, then **link stream**

Stream graphs

$S = (T, V, W, E)$

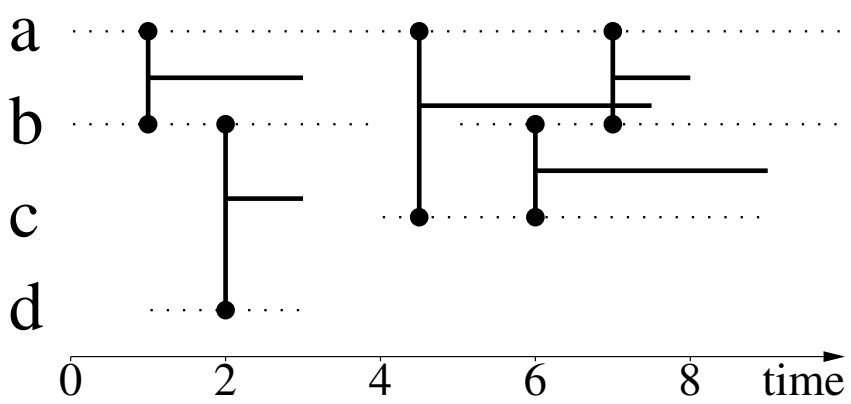
$T = [\alpha, \omega]$ $\quad V = \{u\}$ $\quad W = \{(t, v)\}$ $\quad E = \{(t, uv)\}$



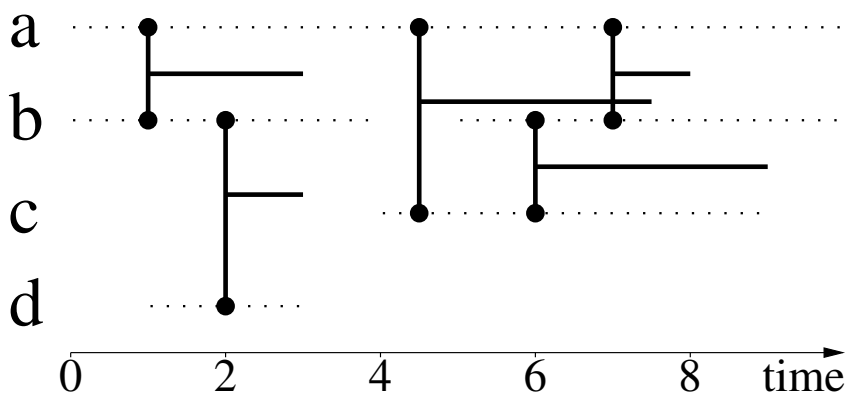
c is present from time 4 to time 9: $\{c\} \times [4, 9] \in W$
a interacts with b from time 1 to time 3: $\{ab\} \times [1, 3] \in E$
...

T_u = presence of node u
$\rightarrow T_b = [0, 4] \cup [4.5, 10]$
T_{uv} = presence of link uv
$\rightarrow T_{ab} = [1, 3] \cup [7, 8]$
If $\forall v, T_v = T$ and $\forall u, v, T_{uv} = \{\emptyset, T\}$, then **graph-equivalent**

Size, duration and compactness



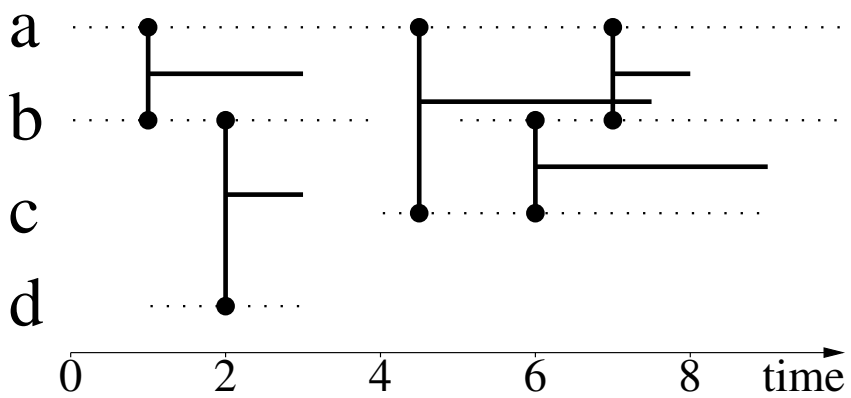
Size, duration and compactness



$$n = \sum_v \frac{|T_v|}{|T|} = \frac{|W|}{|T|}$$

$1 + 0.9 + 0.5 + 0.2 = 2.6$ nodes

Size, duration and compactness



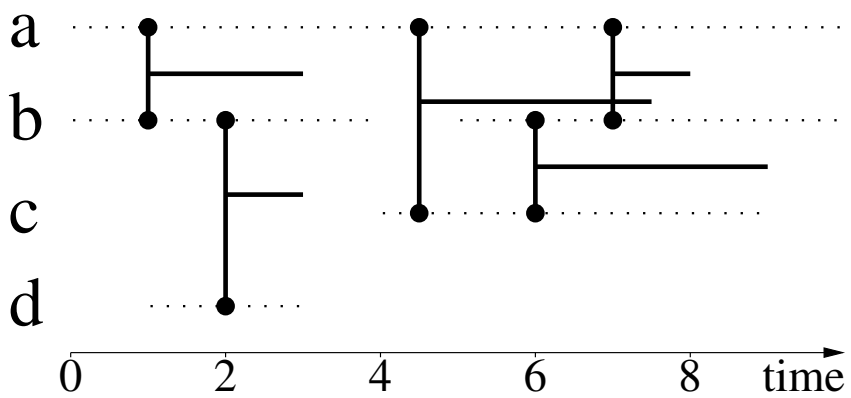
$$n = \sum_v \frac{|T_v|}{|T|} = \frac{|W|}{|T|}$$

$1 + 0.9 + 0.5 + 0.2 = 2.6$ nodes

$$m = \sum_v \frac{|T_{uv}|}{|T|} = \frac{|E|}{|T|}$$

$0.3 + 0.3 + 0.3 + 0.1 = 1$ link

Size, duration and compactness



$$n = \sum_v \frac{|T_v|}{|T|} = \frac{|W|}{|T|}$$

$1 + 0.9 + 0.5 + 0.2 = 2.6$ nodes

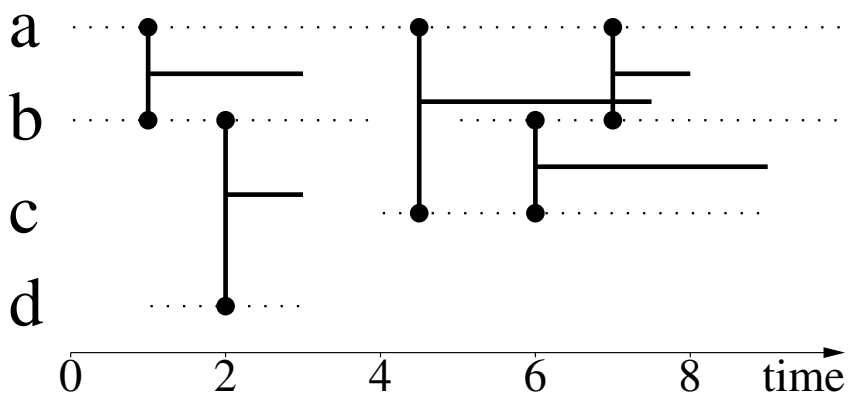
$$m = \sum_v \frac{|T_{uv}|}{|T|} = \frac{|E|}{|T|}$$

$0.3 + 0.3 + 0.3 + 0.1 = 1$ link

$$k = \int_{t \in T} \frac{|V_t|}{|V|} dt = \frac{|W|}{|V|}$$

$\frac{26}{4} = 6.5$ time units

Size, duration and compactness



$$n = \sum_v \frac{|T_v|}{|T|} = \frac{|W|}{|T|}$$

$1 + 0.9 + 0.5 + 0.2 = 2.6$ nodes

$$m = \sum_v \frac{|T_{uv}|}{|T|} = \frac{|E|}{|T|}$$

$0.3 + 0.3 + 0.3 + 0.1 = 1$ link

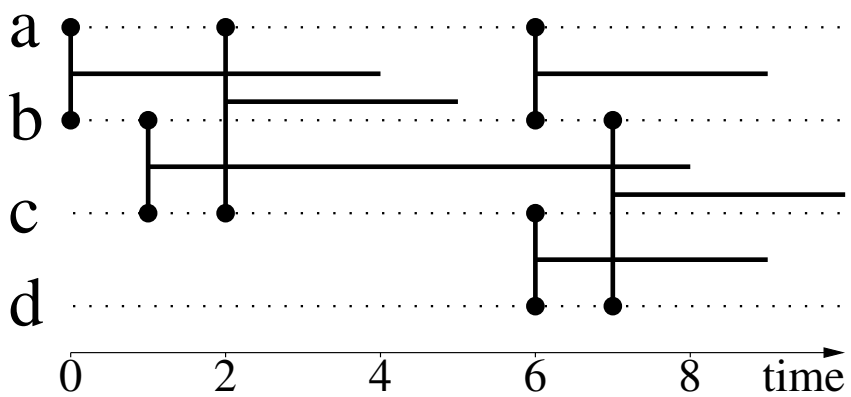
$$k = \int_{t \in T} \frac{|V_t|}{|V|} dt = \frac{|W|}{|V|}$$

$\frac{26}{4} = 6.5$ time units

$$l = \int_{t \in T} \frac{|E_t|}{|V \otimes V|} dt = \frac{|E|}{|V \otimes V|}$$

$\frac{10}{6} \approx 1.66$ time units

Size, duration and compactness



$$n = \sum_v \frac{|T_v|}{|T|} = \frac{|W|}{|T|}$$

$1 + 0.9 + 0.5 + 0.2 = 2.6$ nodes

$$m = \sum_v \frac{|T_{uv}|}{|T|} = \frac{|E|}{|T|}$$

$0.3 + 0.3 + 0.3 + 0.1 = 1$ link

$$k = \int_{t \in T} \frac{|V_t|}{|V|} dt = \frac{|W|}{|V|}$$

$\frac{26}{4} = 6.5$ time units

$$l = \int_{t \in T} \frac{|E_t|}{|V \otimes V|} dt = \frac{|E|}{|V \otimes V|}$$

$\frac{10}{6} \approx 1.66$ time units

S is **compact** iff $\forall v \in V, T_v = [b, e] \subseteq T$

Substreams and clusters

Graphs:

Subgraph $= G' = (V', E')$ such that $V' \subseteq V, E' \subseteq E$

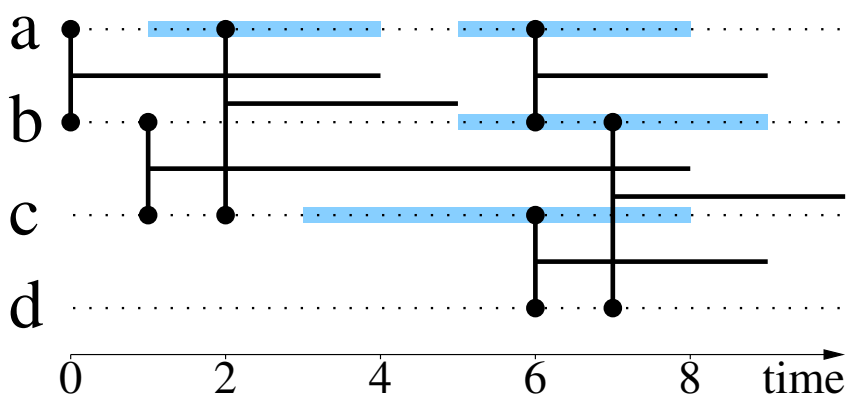
Cluster of nodes $C =$ set of nodes

Stream graphs:

Substream $= S' = (T', V', W', E')$ such that

$T' \subseteq T, V' \subseteq V, W' \subseteq W, E' \subseteq E$

Cluster of nodes $C =$ set of (t, v)



Density

Graphs :

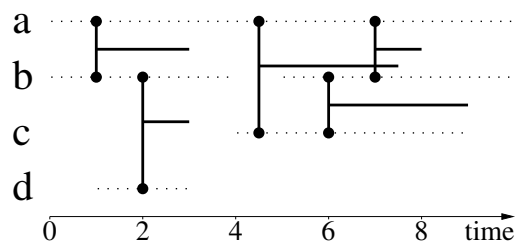
u, v random $=$ link?

$$\delta(G) = \frac{2m}{n \cdot (n-1)}$$

Stream graphs:

uv, t random $=$ link?

$$\delta(S) = \frac{\sum_{uv \in V \otimes V} T_{uv}}{\sum_{uv \in V \otimes V} T_u \cap T_v}$$



Graph-equivalent streams: $\delta(S) = \delta(G)$

Density

Graphs :

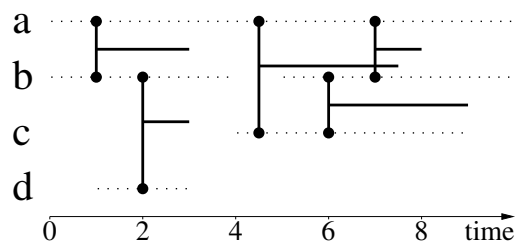
u, v random $=$ link?

$$\delta(G) = \frac{2m}{n \cdot (n-1)}$$

Stream graphs:

uv, t random $=$ link?

$$\delta(S) = \frac{\sum_{uv \in V \otimes V} T_{uv}}{\sum_{uv \in V \otimes V} T_u \cap T_v}$$



$$\delta(S) = \frac{10}{22} \approx 0.45$$

Graph-equivalent streams: $\delta(S) = \delta(G)$

Neighborhood

Graphs :

Neighborhood = set of nodes

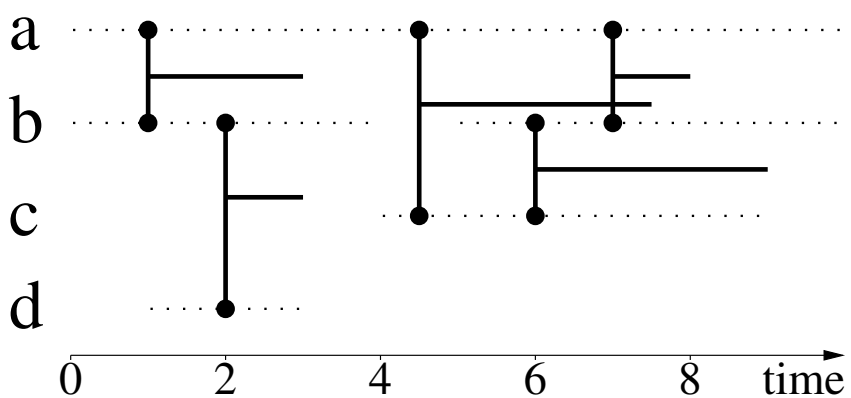
$N(u) = \{v : uv \in E\}$ $d(u) = |N(u)|$

Stream graphs:

Neighborhood = cluster

$N(u) = \{(t, v) : (t, uv) \in E\}$

$d(u) = \frac{|N(u)|}{|T|}$



Graphs, stream graphs: $\sum_u d(u) = 2m$

Neighborhood

Graphs :

Neighborhood = set of nodes

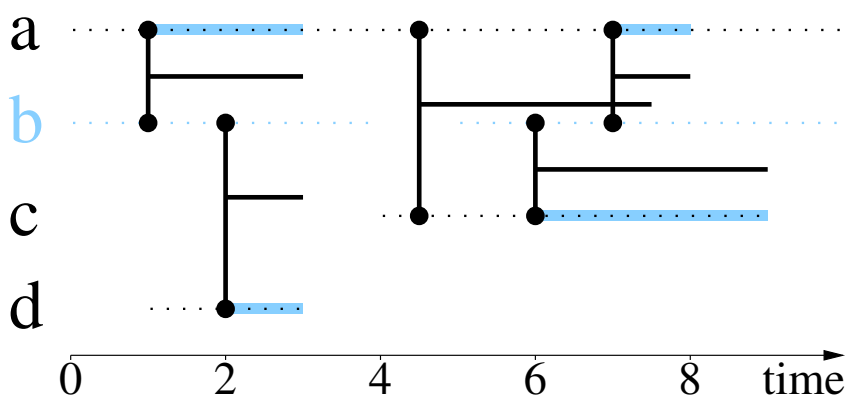
$N(u) = \{v : uv \in E\}$ $d(u) = |N(u)|$

Stream graphs:

Neighborhood = cluster

$N(u) = \{(t, v) : (t, uv) \in E\}$

$d(u) = \frac{|N(u)|}{|T|}$



Graphs, stream graphs: $\sum_u d(u) = 2m$

Degrees

$$d(v) = \frac{|N(v)|}{|T|} = \int_t \frac{d_t(v)}{|T|} dt \qquad d(t) = \sum_v \frac{d_t(v)}{|V|}$$

$$\hat{d}(v) = \int_t \frac{d_t(v)}{|T_v|} dt \qquad \hat{d}(t) = \sum_v \frac{d_t(v)}{|V_t|}$$

$$d(V) = \sum_{v \in V} \frac{|T_v|}{|W|} d(v) \qquad d(T) = \int_t \frac{|V_t|}{|W|} d(t) dt$$

$$d(S) = \sum_v \frac{1}{|V|} d(v) = \frac{2 \cdot |E|}{|T \times V|} = \int_t \frac{1}{|T|} d(t) dt$$

$$\hat{d}(S) = \frac{\sum_v \int_t d_t(v) dt}{|W|} = \frac{2 \cdot |E|}{|W|} = \frac{2m}{n}$$

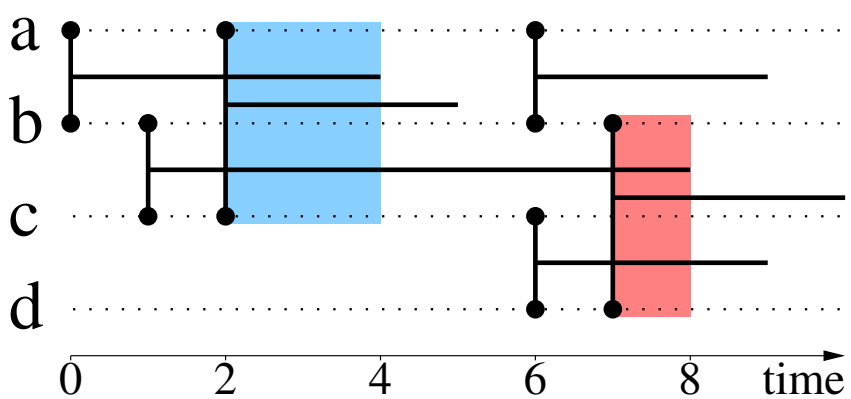
Link stream

$\implies d(v) = \hat{d}(v), d(t) = \hat{d}(t), d(V) = d(T) = d(S) = \hat{d}(S)$

Cliques

Graphs:
Clique X = subgraph $G(X)$ with density 1
Max. if included in no other clique

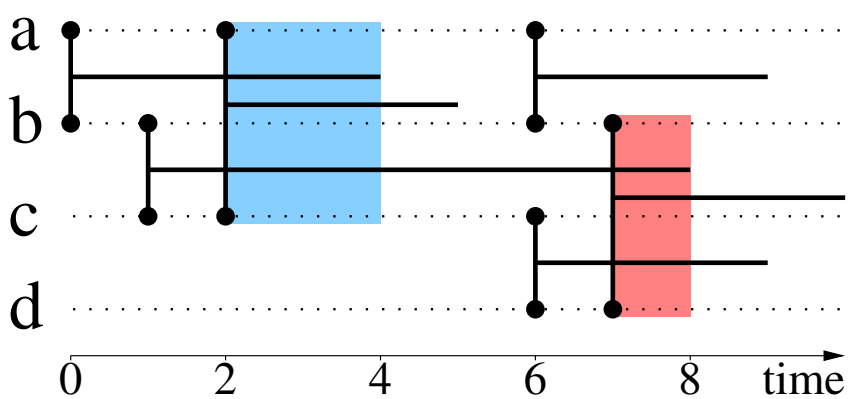
Stream graphs:
Clique X = cluster with density 1
Max. if included in no other clique



Cliques

Graphs:
Clique X = subgraph $G(X)$ with density 1
Max. if included in no other clique

Stream graphs:
Clique X = cluster with density 1
Max. if included in no other clique



Computing maximal compact cliques of S in $\mathcal{O}(2^n n^2 m^3 + 2^n n^3 m^2)$ time, and $\mathcal{O}(2^n n m^2)$ space

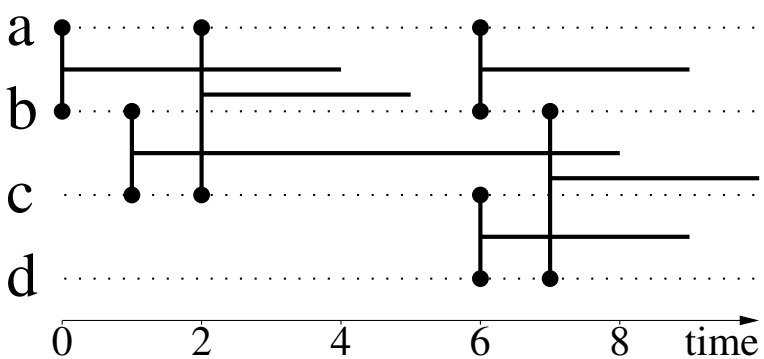
Clustering coefficient

Graphs:

$$cc(u) = \delta(N(u))$$

Stream graphs:

$$cc(u) = \delta(N(u))$$



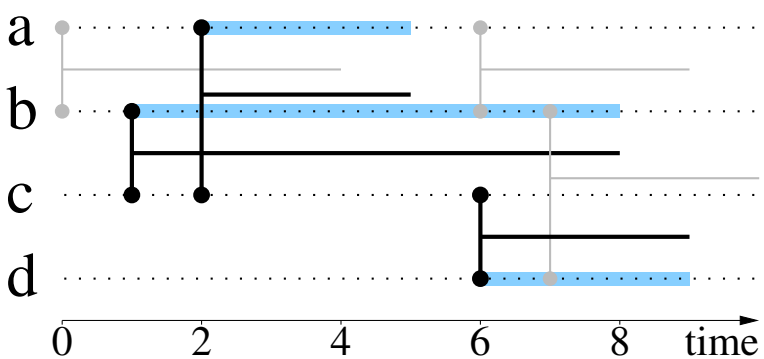
Clustering coefficient

Graphs:

$$cc(u) = \delta(N(u))$$

Stream graphs:

$$cc(u) = \delta(N(u))$$



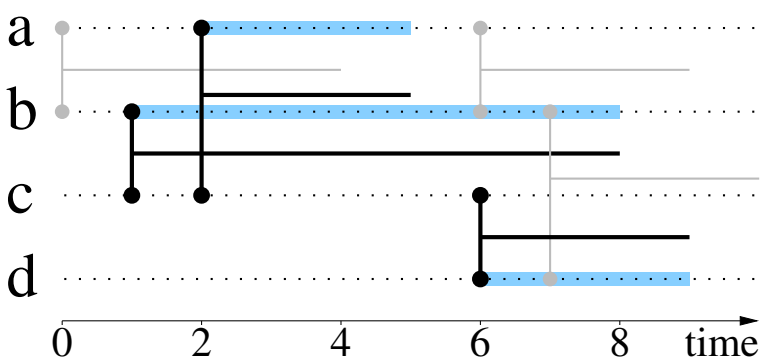
Clustering coefficient

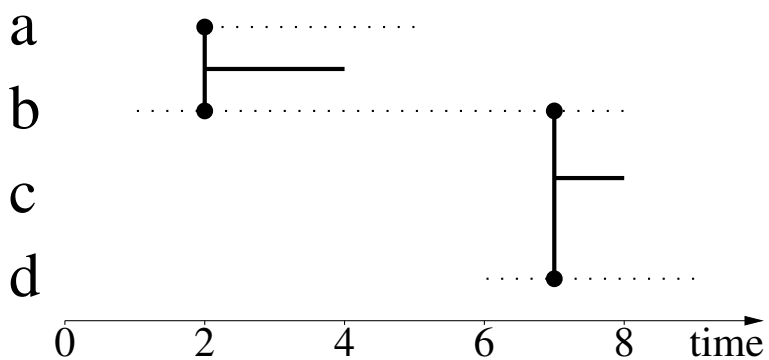
Graphs:

$$cc(u) = \delta(N(u))$$

Stream graphs:

$$cc(u) = \delta(N(u))$$





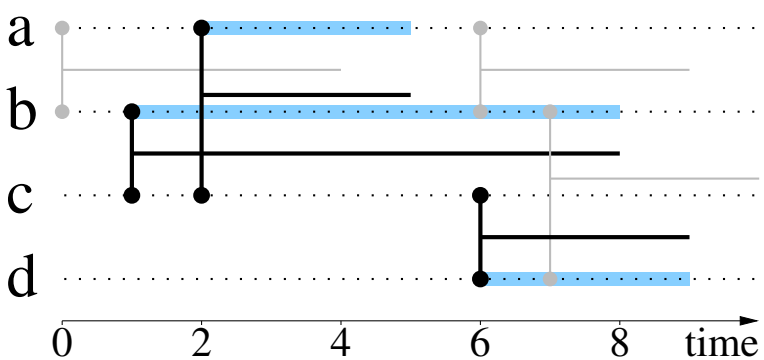
Clustering coefficient

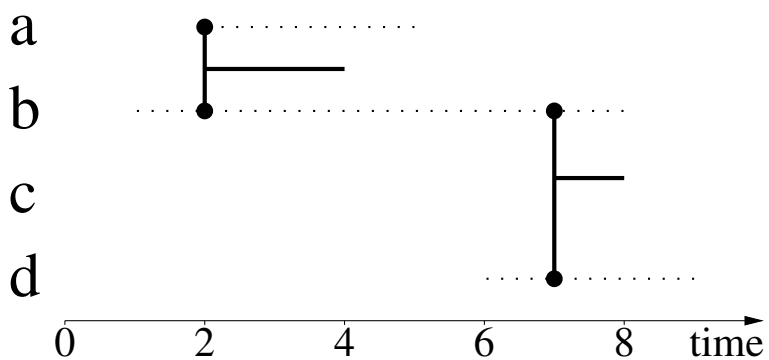
Graphs:

$$cc(u) = \delta(N(u))$$

Stream graphs:

$$cc(u) = \delta(N(u))$$





$$cc(c) = \frac{3}{5} = 0.6$$

Outline

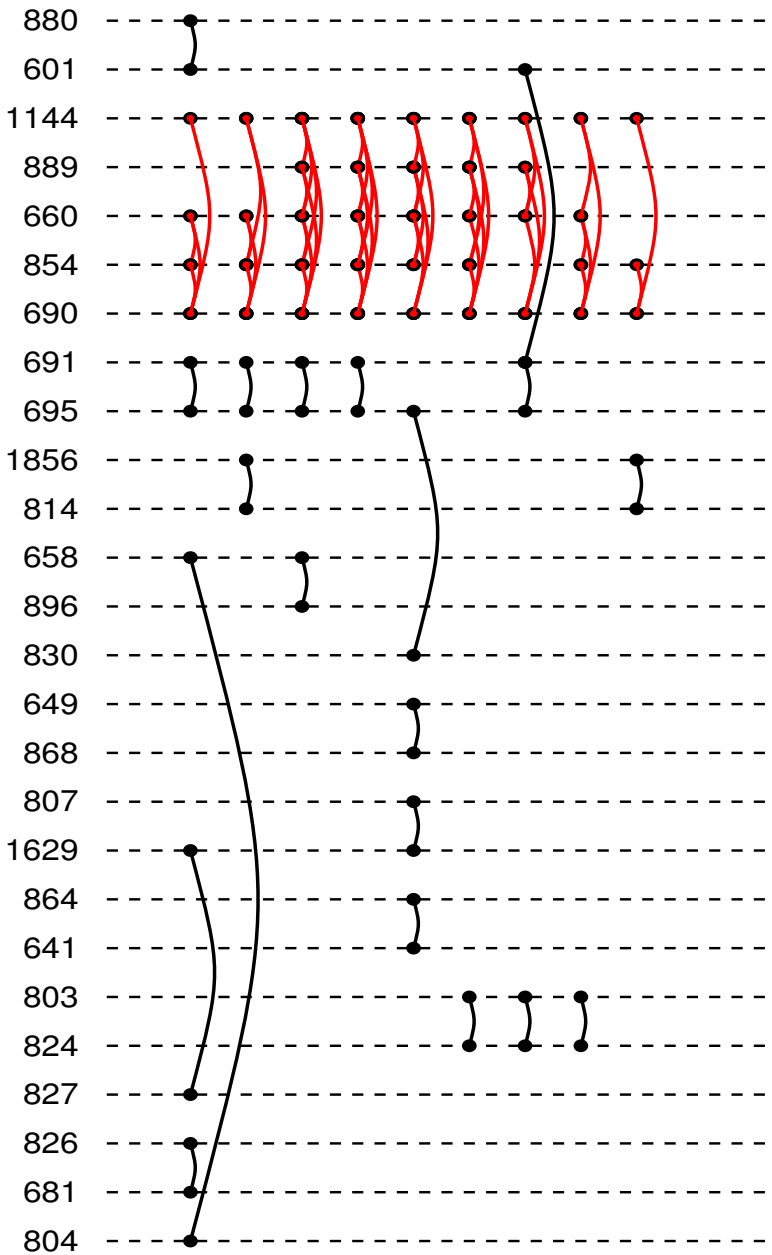
Formalism

- **Stream graphs**
- **Clusters**
- **Density**
  - **Neighborhood** and degrees
  - **Cliques**
  - **Clustering coefficient**, transitivity
  - Quotient stream
  - k-cores
- Paths
  - Accessibility
  - Connectedness
  - Centralities
  - Trees and cascades

Data analysis

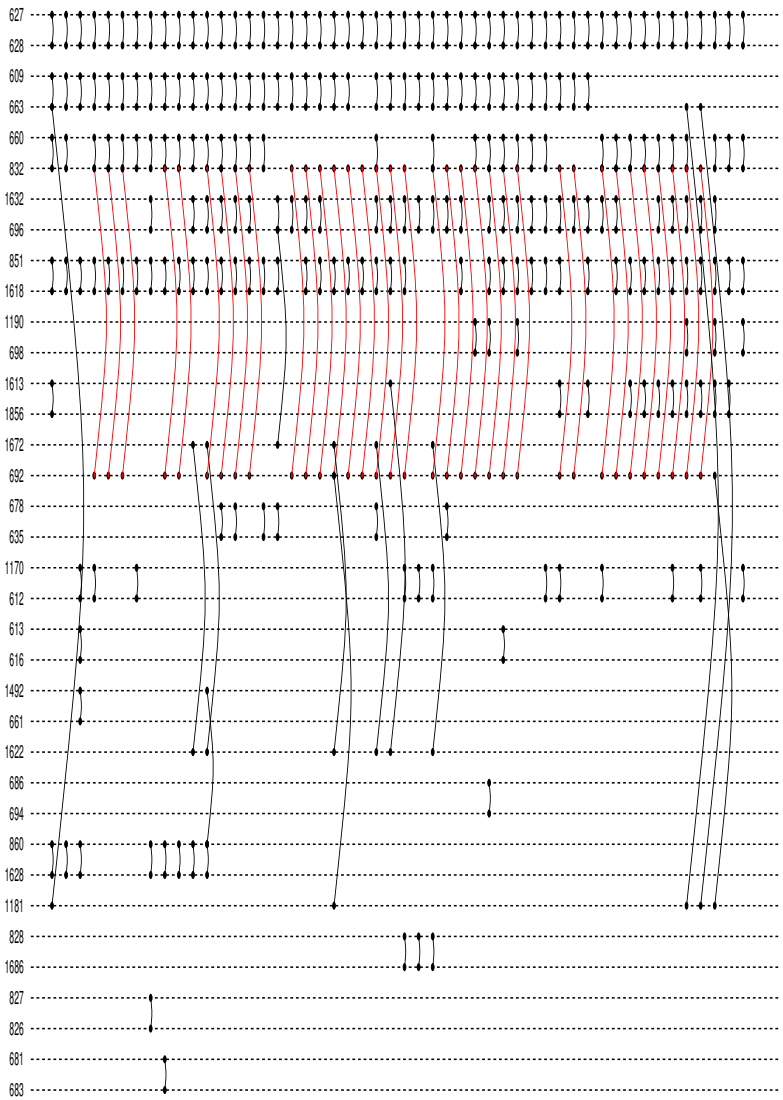
- **Face-to-face contacts**
  - **cliques**
  - **visualization**
- **Emails**
  - **threads = communities?**
- IP traffic
  - Statistical analysis
  - Bipartite stream graphs
  - Cliques for anomaly detection
- Machine Learning
  - feature engineering
  - rating prediction

Results on a real-world contact trace



- Contacts between high-school students (SocioPatterns)
- 4 granularities: 1mn, 15mn, 1 hour, 3 hours
- A few thousand Δ-cliques (1742 in graph)
- Size up to 7 nodes
- Duration up to 20 hours
- Typically **short meetings**, **work groups**, **lunches**...
- **Not** new patterns, but hard to see with other methods

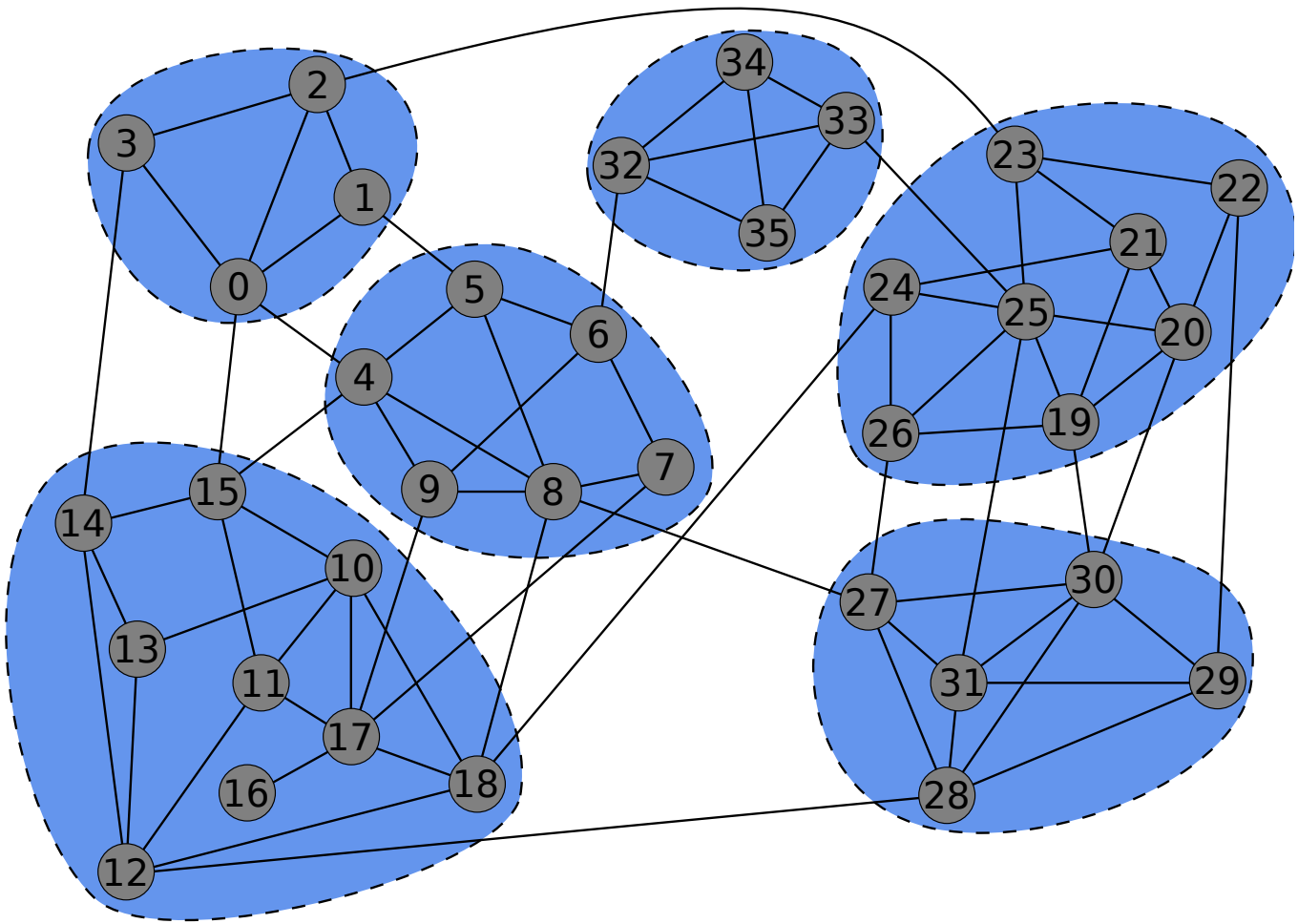
Results on a real-world contact trace



- Contacts between high-school students (SocioPatterns)
- 4 granularities: 1mn, 15mn, 1 hour, 3 hours
- A few thousand Δ-cliques (1742 in graph)
- Size up to 7 nodes
- Duration up to 20 hours
- Typically **short meetings**, **work groups**, **lunches**...
- **Not** new patterns, but hard to see with other methods

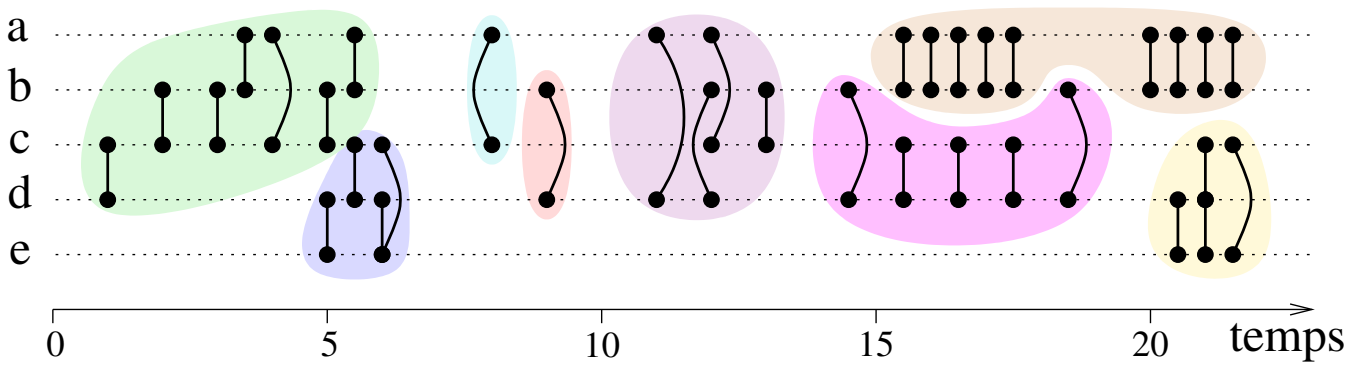
Heading to communities?

Community = dense subgraphs sparsely interconnected



Heading to communities?

Community = dense substreams sparsely interconnected?



Dataset

Debian mailing-list archive (`debian-user`)

Contains **thread information** (IN-REPLY-TO)

722,716 messages involving 51,753 authors

Most threads: ≥ 3 messages, largest : 100 messages

20 years of data (*01/1996 – 12/2014*)

Most threads: a few days, longest : 1 year

Maximal cliques = arguments

Inter- and intra- thread densities

$\mathcal{C} = \{C_i\}_{i=0}^{k}$ a partition of the stream graph

Intra-cluster density

$$t, u, v \text{ random}, \ u, v \in V(C_i), \ t \in T(C_i) : (t, uv) \in E?$$

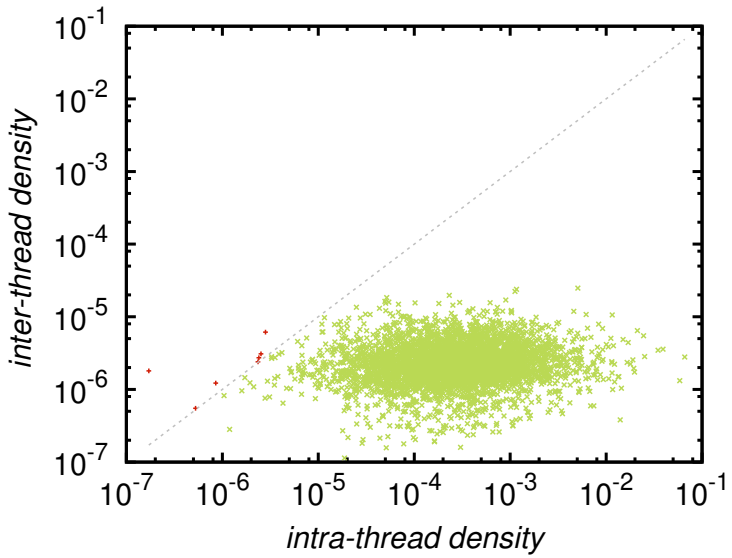
Inter-cluster density

$$t, u, v \text{ random}, \ u \in V(C_i), v \in V(C_j), \ t \in T(C_i \cup C_j) : \\ (t, uv) \in E?$$

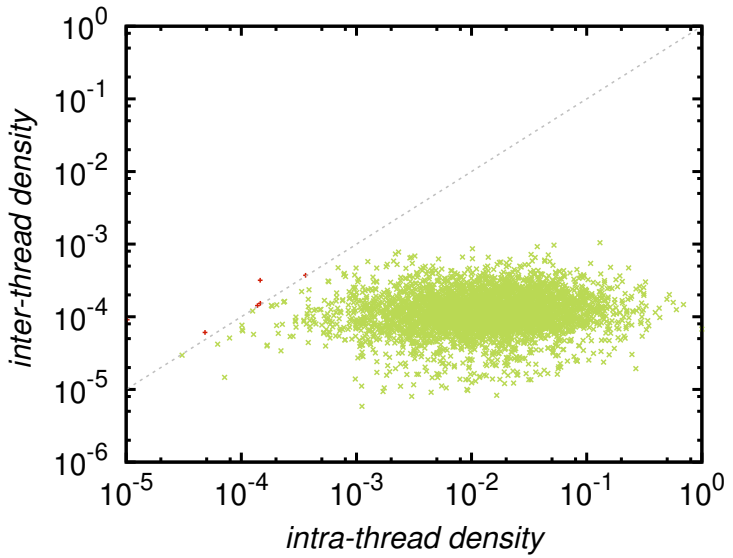
Study at 4 granularities:

- 1 minute
- 1 day
- 1 year
- 20 years

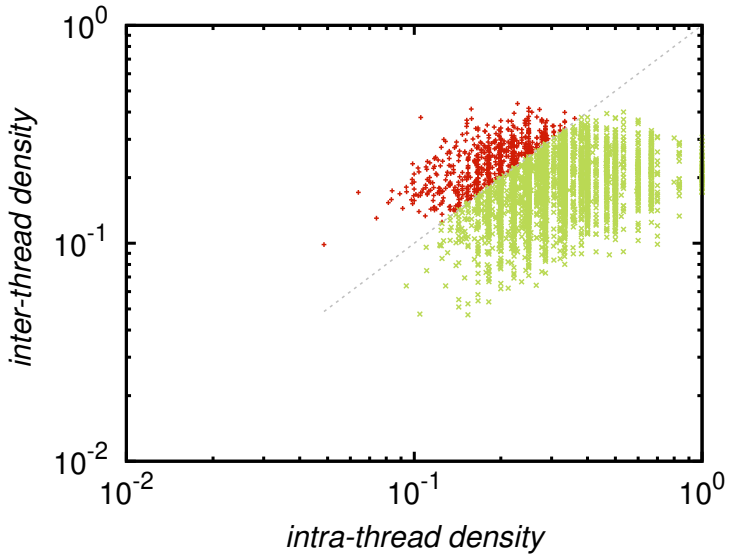
Inter- and intra- thread densities correlations



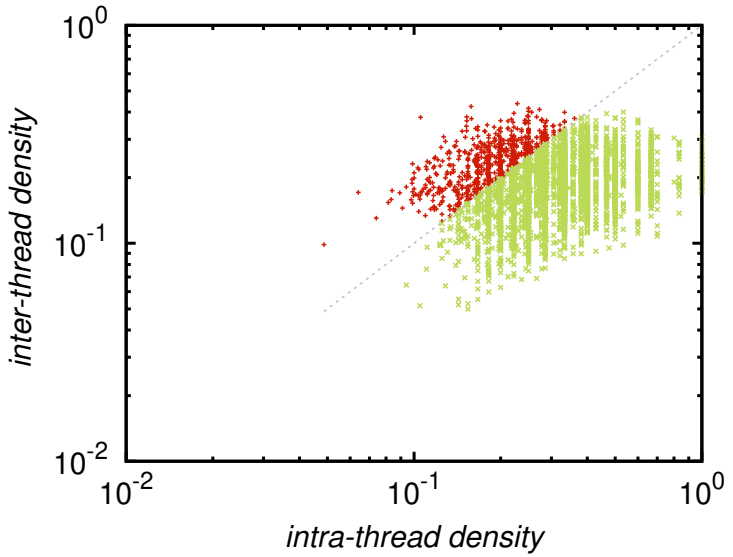
Δ =1 minute



Δ =1 hour



Δ =1 year



Δ =20 years

Wrapping up

1. A versatile language for interaction streams

   $\rightarrow$ Theoretical and algorithmic development

2. Application to numerous real-world examples

   $\rightarrow$ Link prediction, random generators, community detection

[1] Stream graphs and link streams for the modelling of interactions over time. Matthieu Latapy, Tiphaine Viard, Clémence Magnien. arxiv.org/pdf/1710.04073.pdf

Thank you for your attention