TENSOR -BASED METHODS FOR TEMPORAL NETWORKS

Laetitia Gauvin
In collaboration with Anna Sapienza, Ciro Cattuto
Alain Barrat, André Panisson

Machine learning in network science

NetSci

INSTITUTE
FOR SCIENTIFIC INTERCHANGE
FOUNDATION

2

CONTEXT

NETWORKS

- Social (facebook, Twitter)
- Infrastructure (transportation)
- Communication (emails, phone)

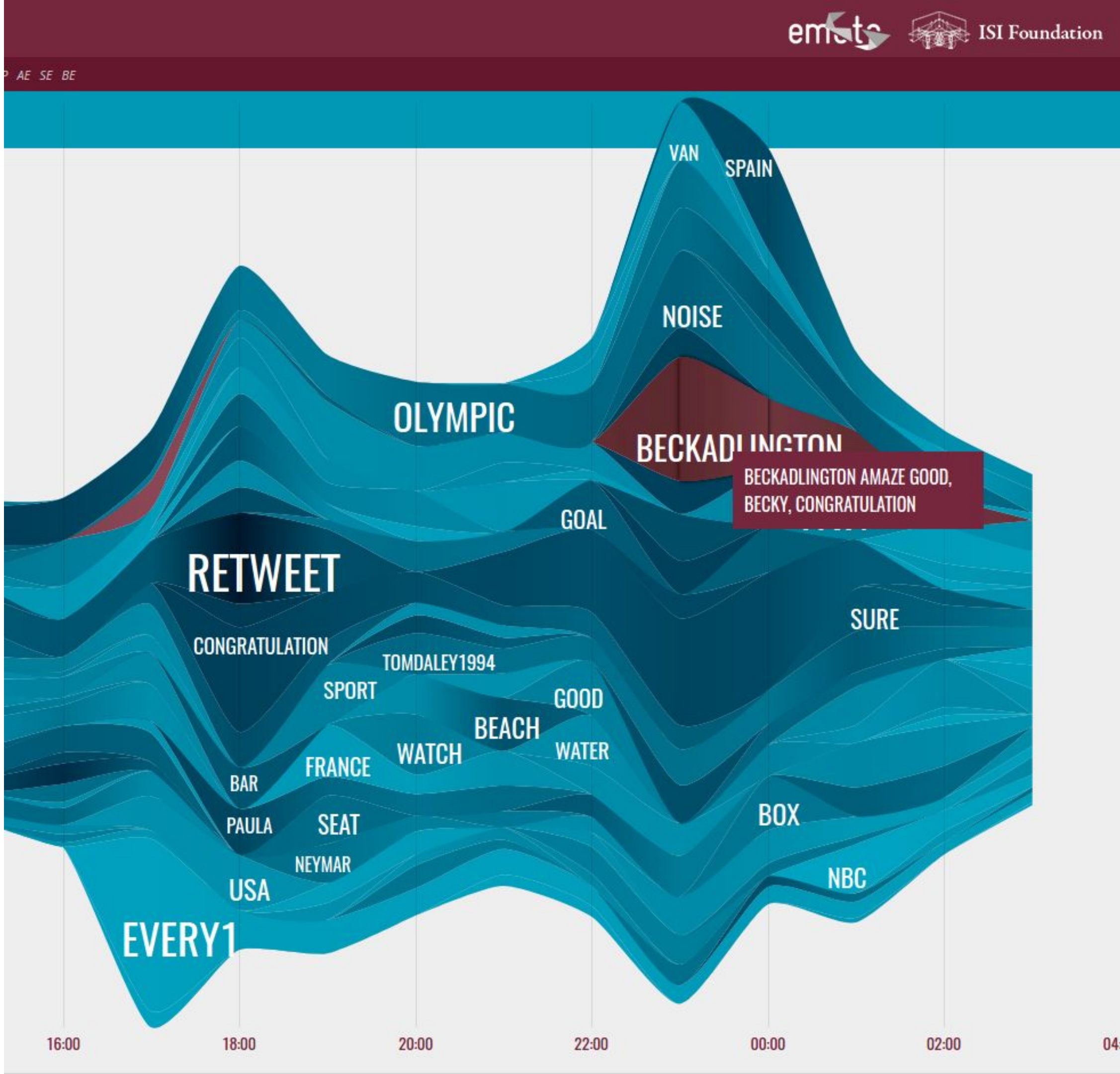
DIMENSIONS

- temporal
- structural
- spatial

How to capture the different properties of networks relevant for complex phenomena?

DIMENSIONALITY REDUCTION

Transformation of data into a meaningful representation of reduced dimension



A Panisson, L Gauvin, M Quaggiotto, C Cattuto, Mining Concurrent Topical Activity in Microblog Streams, Proceedings of the the 4th Workshop on Making Sense of Microposts co-located with the 23rd International World Wide Web Conference (WWW 2014)

OUTLINE

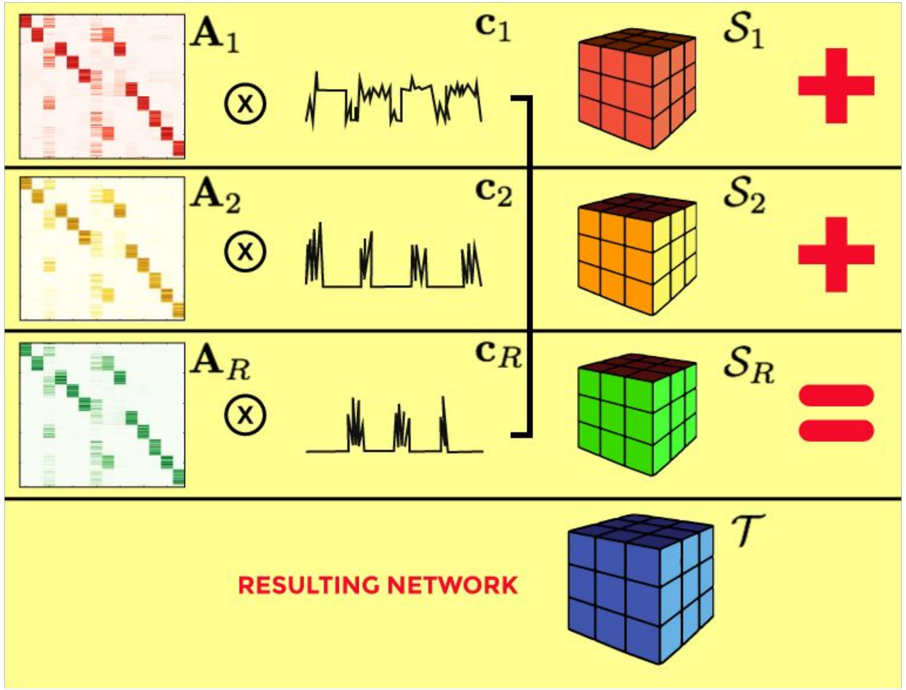
1) Structure discovery

2) Structures & spreading processes

3) Structure recovery & spreading processes

1.
STRUCTURE DISCOVERY

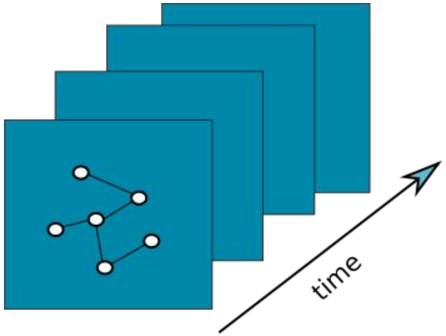
"NETWORK AS A SUM OF NETWORKS"

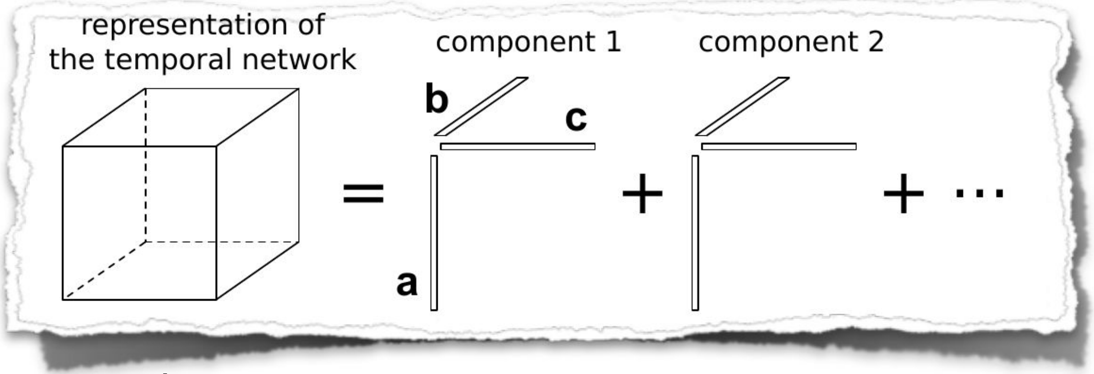


$$\mathcal{T} = \sum_{r=1}^{R} \mathbf{A}_r \circ \mathbf{c}_r$$

$$\mathbf{A}_r = \mathbf{a}_r \circ \mathbf{b}_r$$

DETECTION OF MESOSCALE STRUCTURES





R mesoscale structures

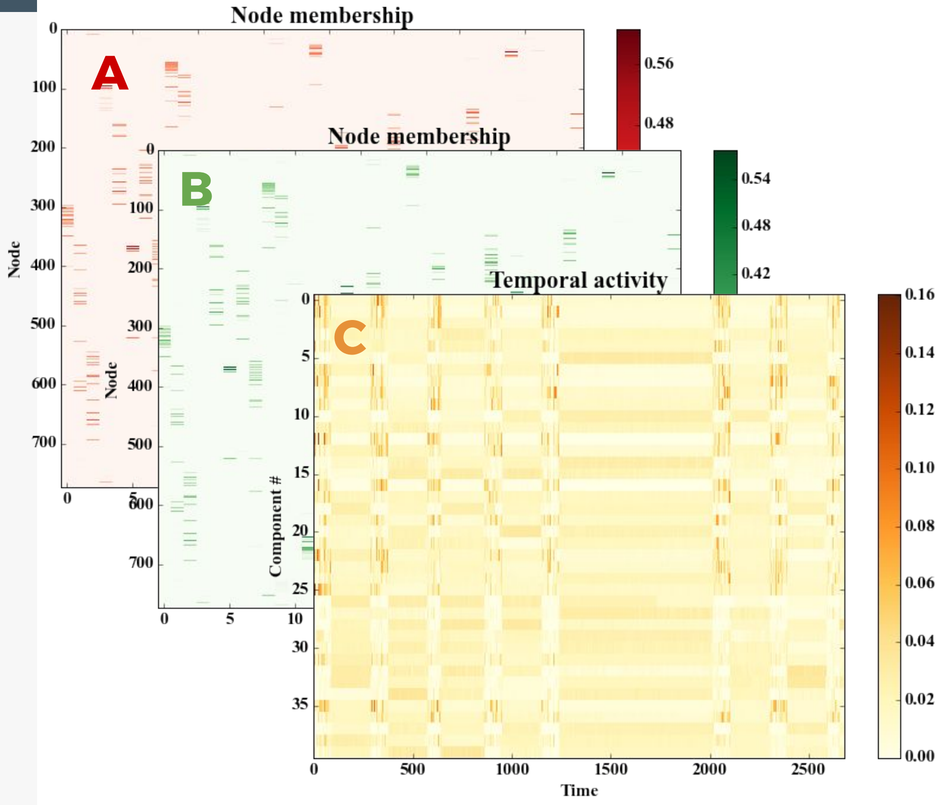
$$\mathcal{T} = \sum_{r=1}^{R_{\mathcal{T}}} \mathbf{a_r} \circ \mathbf{b_r} \circ \mathbf{c_r}$$

link participation

temporal activity

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications.*SIAM review*, *51*(3), 455-500.

FACTORIZATION OUTPUT

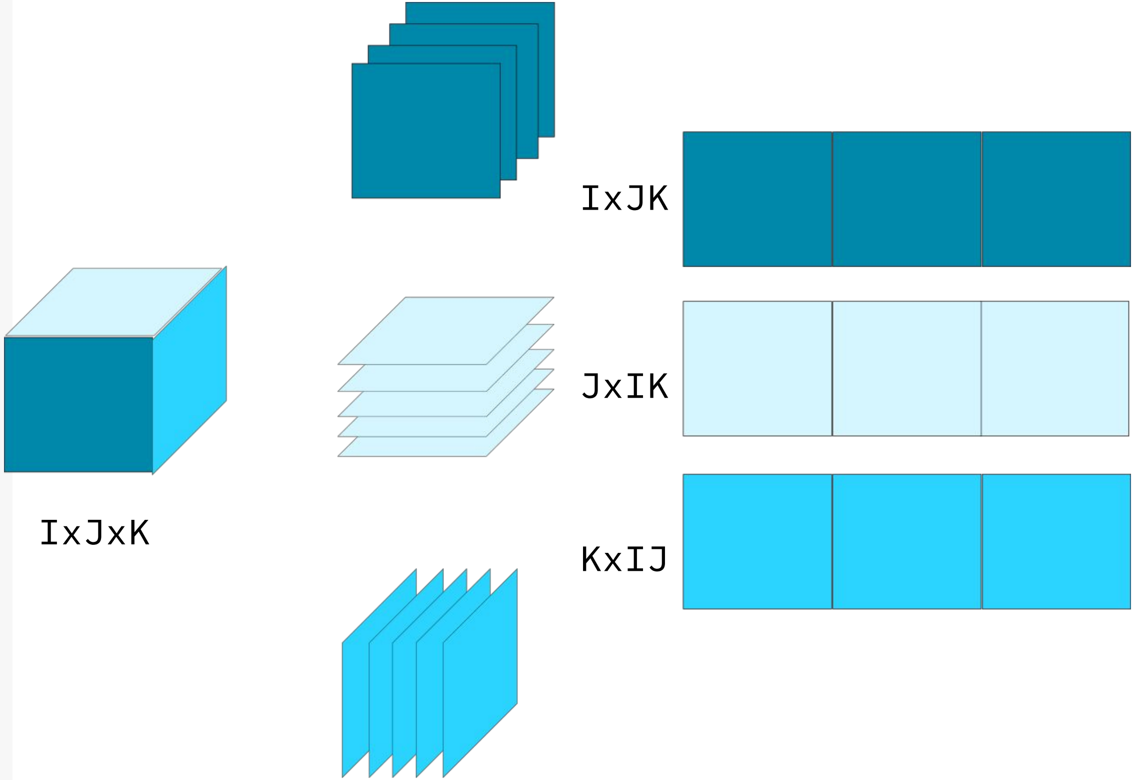


- membership of **nodes** to the components
- membership of **links** to the components

$$a_r \cdot b_r^T$$

- **temporal activity** of the components

MATRICIZATION



$$min\|\mathbf{T}_{(1)} - \mathbf{A}\left(\mathbf{C} \odot \mathbf{B}\right)^T\|_2$$

$$min\|\mathbf{T}_{(2)} - \mathbf{B}\left(\mathbf{C} \odot \mathbf{A}\right)^T\|_2$$

$$min\|\mathbf{T}_{(3)} - \mathbf{C}\left(\mathbf{B} \odot \mathbf{A}\right)^T\|_2$$

KKT CONDITIONS

$$\|\mathbf{V}\mathbf{X}-\mathbf{W}\|_2$$

$$\mathbf{V}=\left(\mathbf{C}^T\mathbf{C}*\mathbf{A}^T\mathbf{A}\right)\ ,\quad \mathbf{X}=\mathbf{B}^T$$

$$\text{and}\quad \mathbf{W}=\mathbf{\Lambda}\left(\mathbf{C}\odot\mathbf{A}\right)^T\mathbf{T}_{(2)}^T\ ,$$

Karush-Kuhn-Tucker (KKT)

$$f\left(\mathbf{X}\right)=\mathbf{V}^T\mathbf{V}\mathbf{X}-\mathbf{V}^T\mathbf{W}$$

$$f\left(\mathbf{X}\right)\geq 0\ ,\ \nabla f\left(\mathbf{X}\right)^T\mathbf{X}=0\ ,\ \mathbf{X}\geq 0$$

$$\mathbf{X}^T\mathbf{V}^T-\mathbf{W}^T=0$$

Fast Nonnegative Tensor Factorization with an Active-set-like Method.,Jingu Kim and Haesun Park, In High-Performance Scientific Computing: Algorithms and Applications, Springer, pp. 311-326, 2012.

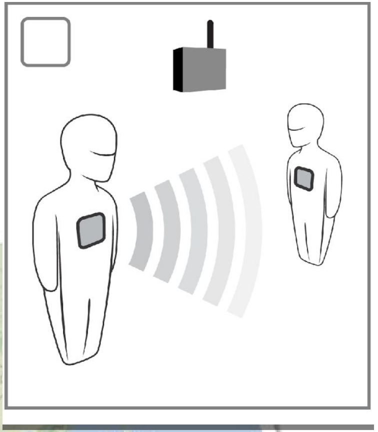
ESTIMATION OF THE NUMBER OF COMPONENTS

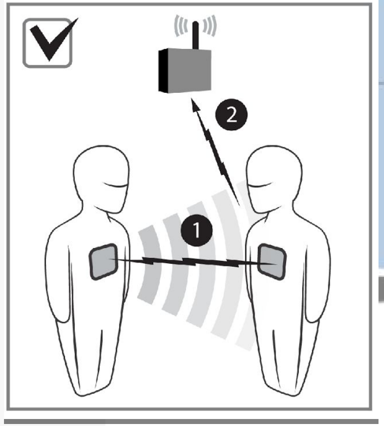
- Core consistency : based on the comparison of the core with Tucker decomposition

- Cophenetic coefficient : based on consensus matrices

Brunet, J. P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences*, *101*(12), 4164-4169.
Bro, R., & Kiers, H. A. (2003). A new efficient method for determining the number of components in PARAFAC models. *Journal of chemometrics*

DATA

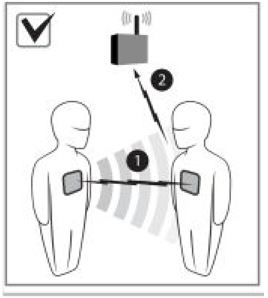


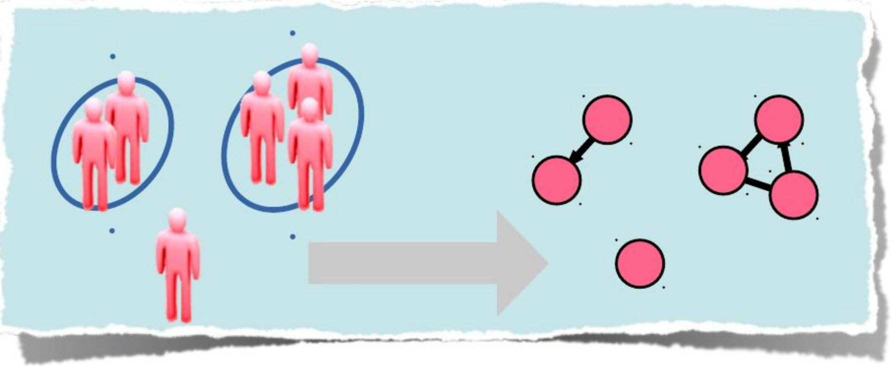


SocioPatterns.org

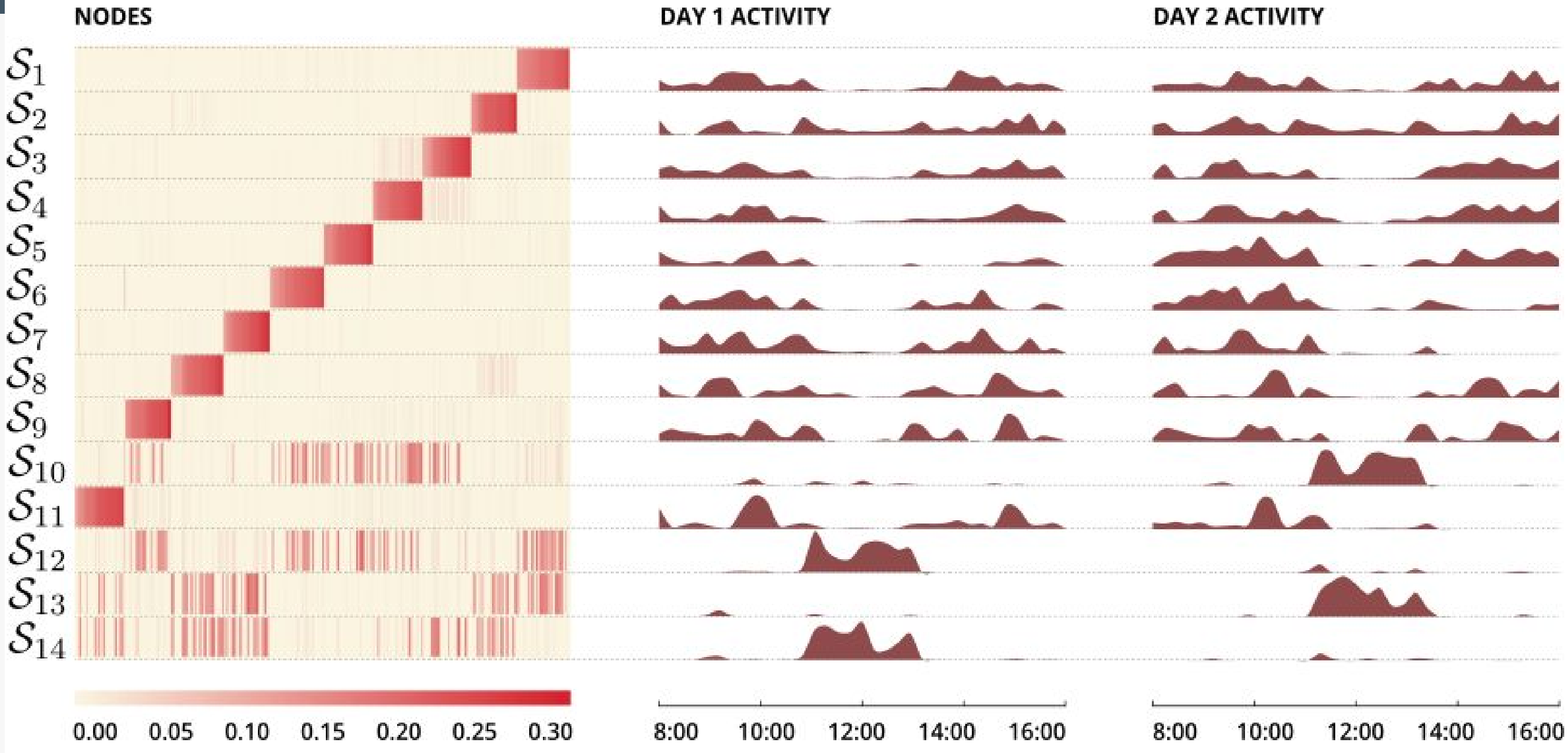
APPLICATION (1)

Lyon,
France
231 students
10 teachers
2 days



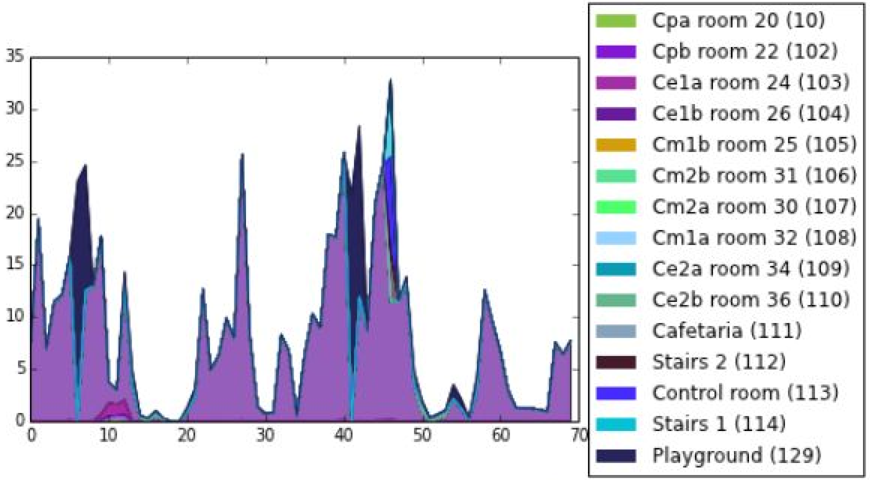


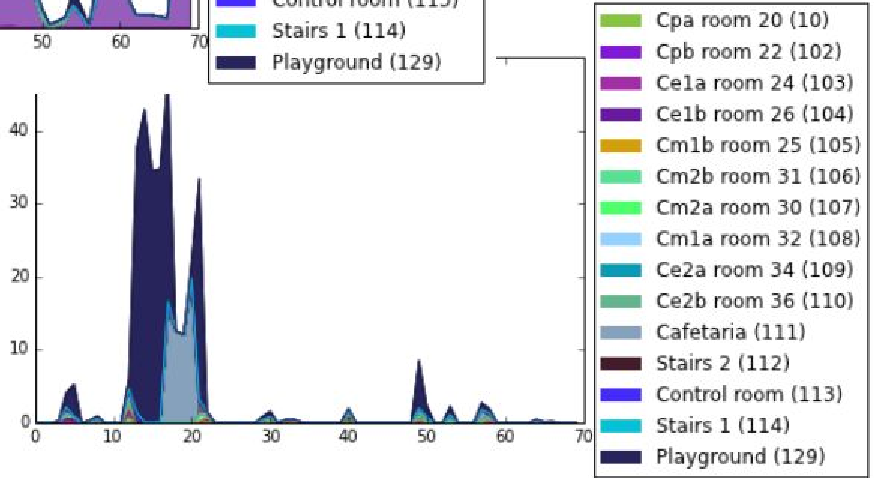
MESOSCALE STRUCTURE DETECTION



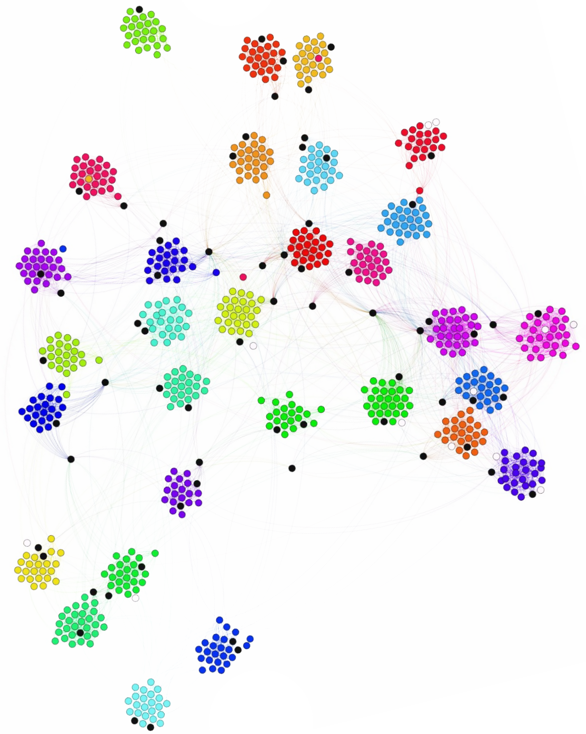
Gauvin, L., et al. Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. *PloS one*, 201

PATTERN DETECTION





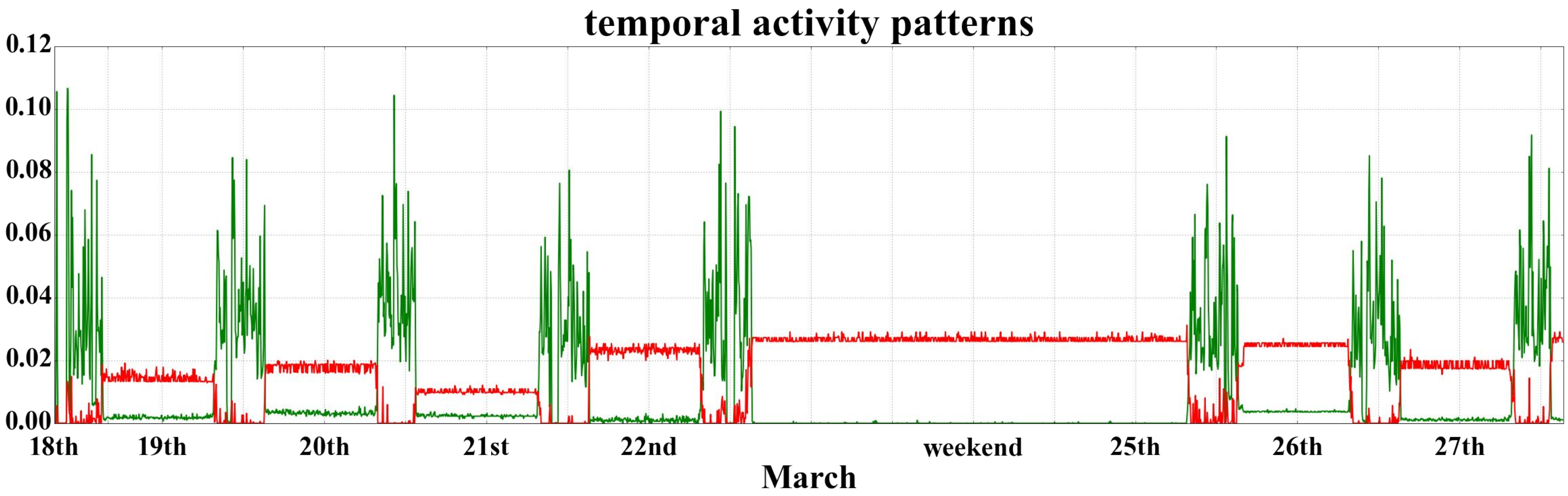
APPLICATION (2)



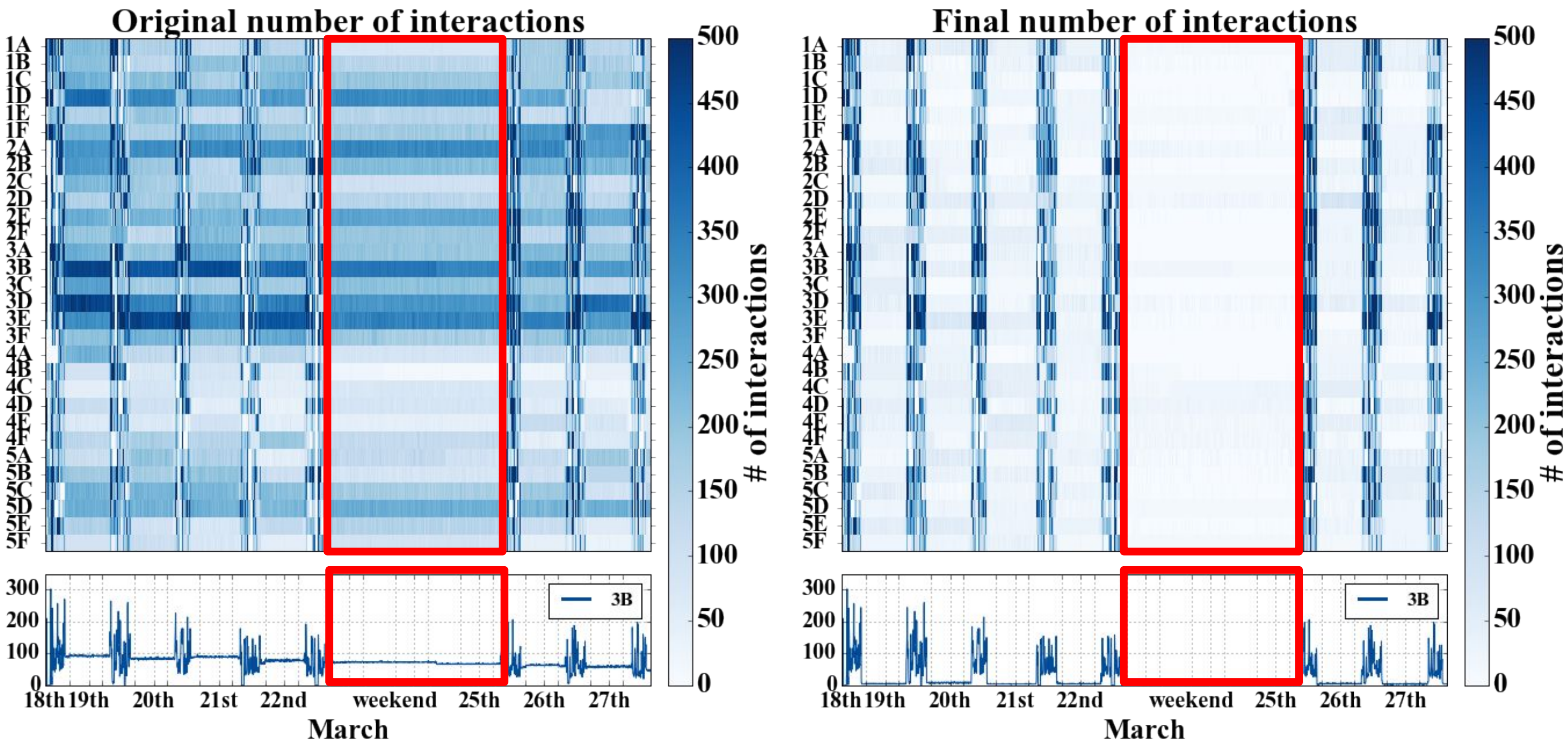
- 709 students
- 65 teachers
- 30 classes

- 10 days
- 5 min resolution

ANOMALY DETECTION



ANOMALY DETECTION

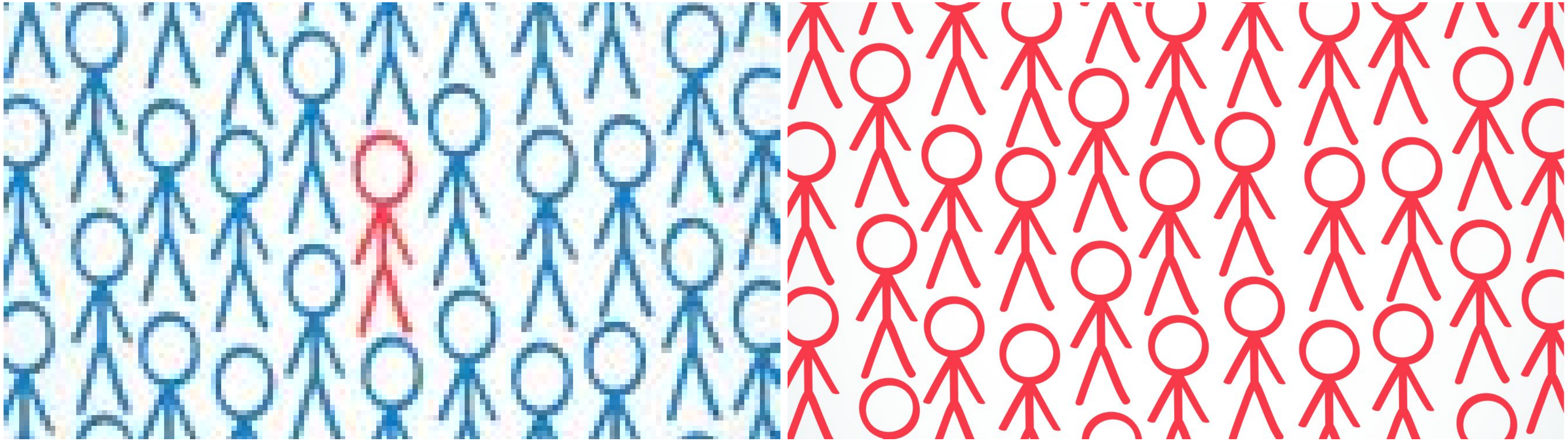


Sapienza, A., et al. "Anomaly Detection in Temporal Graph Data: An Iterative Tensor Decomposition and Masking Approach." *AALTD@ PKDD/ECML*. 2015.

CONCLUSIONS OF PART 1

- Methodology based on non-negative factorization efficient to divide a network in elementary pieces
- Patterns extracted with meaningful interpretation good for tackling several problems encountered in network science

2.
INTERPLAY WITH SPREADING PROCESSES



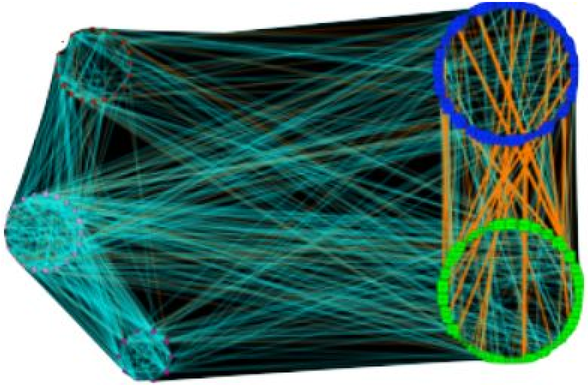
INTERVENTION STRATEGY

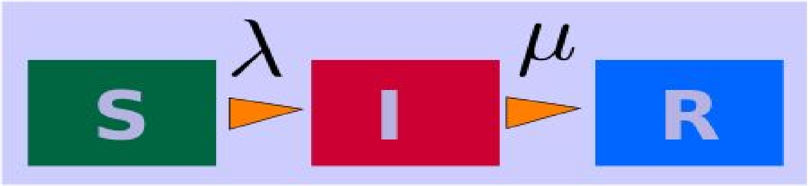
MICROSCOPIC

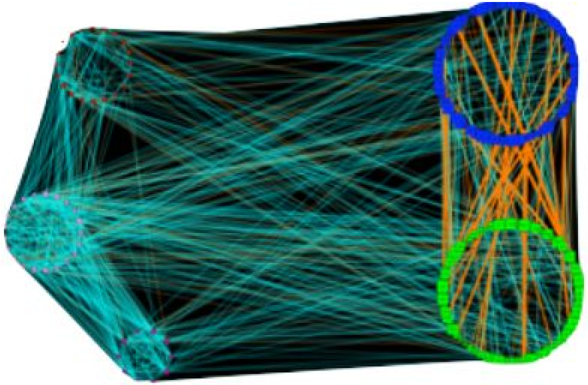
How to mitigate epidemic spread by using both temporal and topological properties of temporal network?

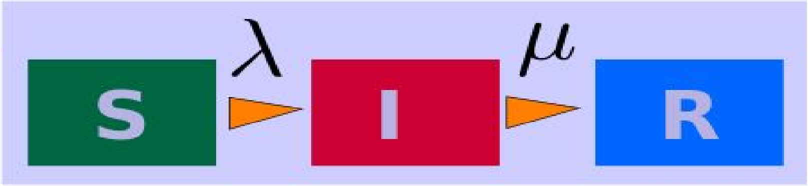
MACROSCOPIC

MESOSCALE TARGETED INTERVENTION: SIR PROCESS



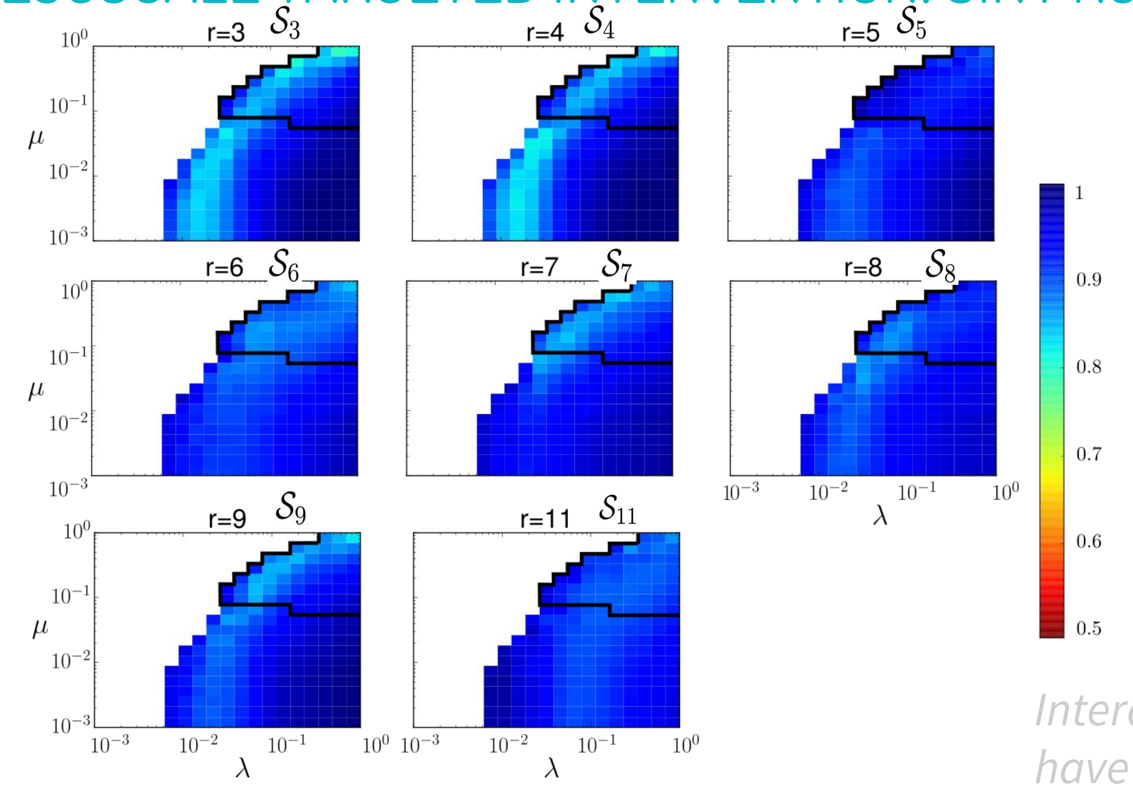






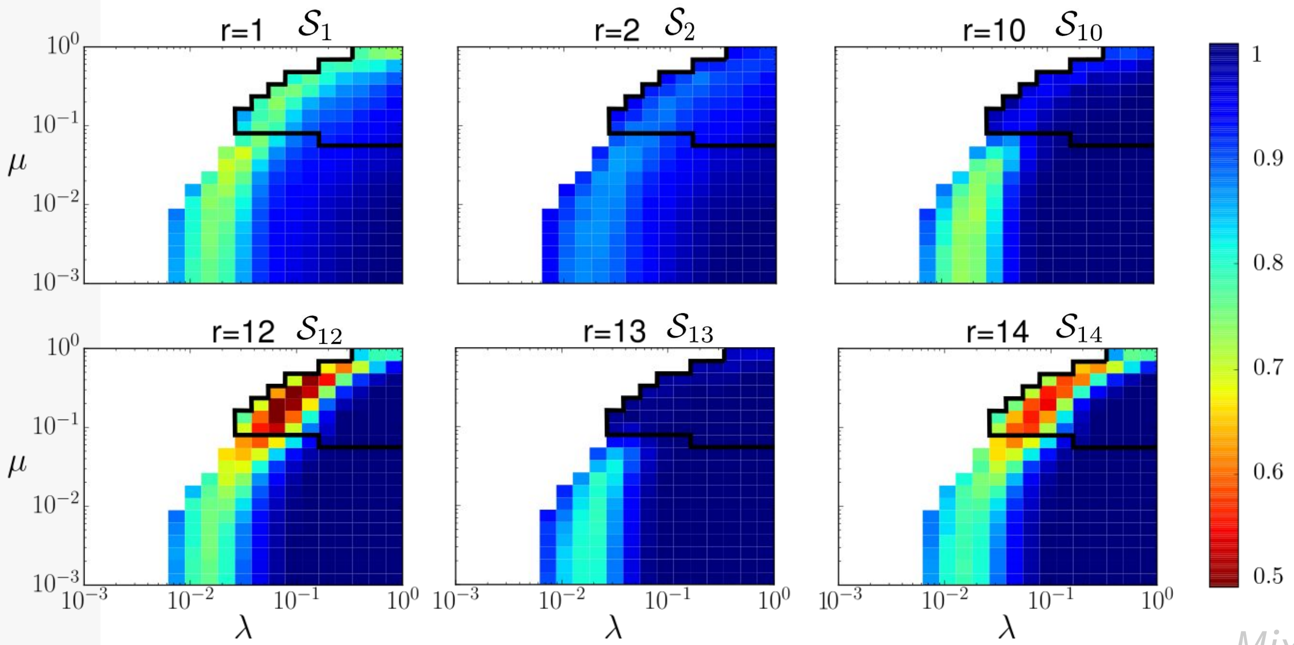
Impact on the epidemic spread

MESOSCALE TARGETED INTERVENTION: SIR PROCESS



Interactions in classes have a very weak role in the spreading process

MESOSCALE TARGETED INTERVENTION: SIR PROCESS



Mixing events are crucial for intervention

Gauvin, L., et al. "Revealing latent factors of temporal networks for mesoscale intervention in epidemic spread." *arXiv preprint arXiv:1501.02758* (2015).

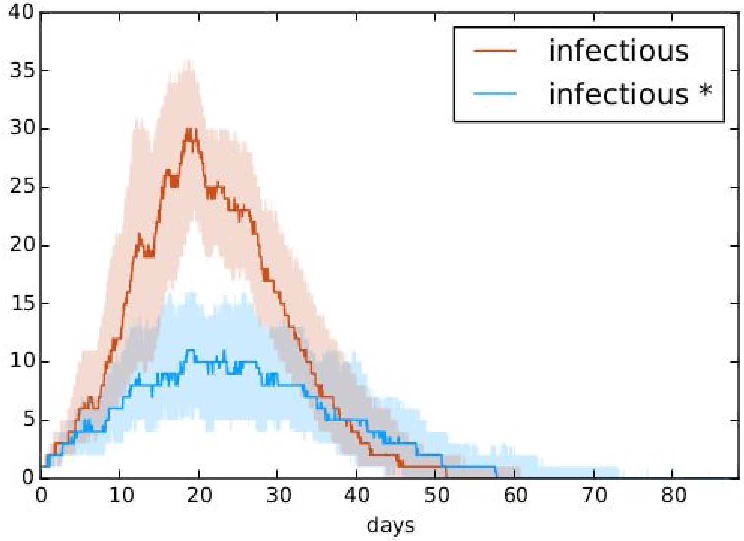
ILI IN A PRIMARY SCHOOL

- Dataset : sequence of typical weeks in the school
- Influenza-like disease : SEIR
- Exposed in the school and outside
- Latent period : 2 days
- Recovery : 4 days
- Infectious go home after school
- Reactive intervention : avoid interactions detected as having a strong impact once the spreading started
- Intervention equivalent to limit mix events and replace by class-like events

ILI IN A PRIMARY SCHOOL:MITIGATION

Percentage of simulations with an attack rate greater than 10%

- 54 % in case of an intervention
- 71 % without intervention



CONCLUSIONS OF PART 2

- Methodology to uncover mesoscale structures in temporal networks in an unsupervised manner and rate their importance in a spreading process
- Targeted intervention :

  no need to involve the whole system

  no need to define a ranking of the nodes
- Non trivial mesostructures but interpretable : complex patterns of correlated activity
- Following the previous framework, we show that a reorganization of the schedule leads to reduction of 42% of infectious cases

3.
MISSING DATA RECOVERY AND SPREADING PROCESSES

MISSING DATA & SPREADING PROCESS

High-resolution interaction data available thanks to social media, electronic devices (RFID, bluetooth…)

MISSING DATA

Lack of participation (i.e. in surveys)

Technical issues during data collection process

...

IMPACT ON SPREADING PROCESS?

Missing data affect temporal and structural properties of contact networks

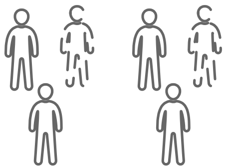
⟹ Inaccurate or misleading results

Main ways to cope with this: ignoring or replacing by mean or statistics

Here we propose an approach at the meso-scale level

IDEA

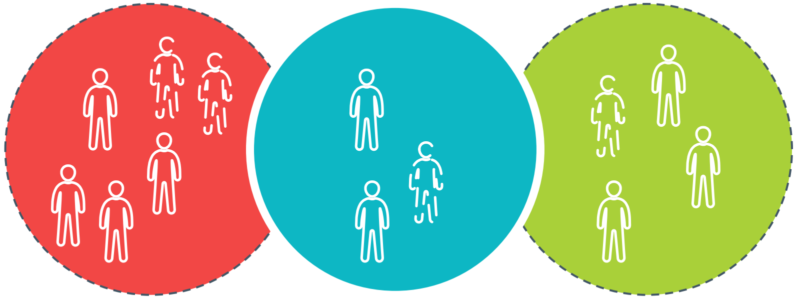
TIME



Nodes with activity partially missing

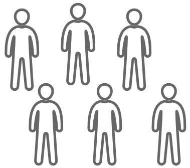
Temporal network with partial information

Extraction of mesoscale structures



Temporal and structural properties

TIME



Use of mesoscale + global properties

Reconstruction of the temporal network

CASE STUDY

Data

Face-to-face contacts (SocioPatterns)

- Conferences

  417 nodes / 3 days

  137 nodes / 2 days

- School

  241 nodes / 2 days

Simulation of missing data

Build the network with partial information

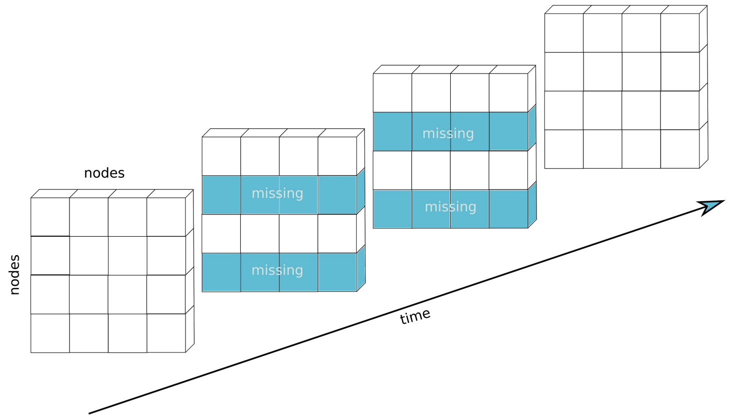
- Select nodes at random & time intervals

  Selection of the links to erase

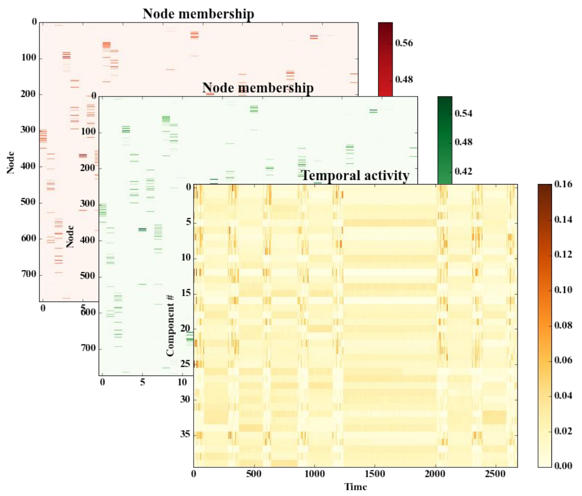
- Imputation of the data accordingly

  Creation of a new network with missing data

RECOVERING MISSING DATA (1)



$$\bar{\mathcal{T}} = \mathcal{T} \boxdot \mathcal{W} + (1 - \mathcal{W})[\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$



Factorization on the partial contact network

“Infer” contact activity of the nodes for which part of the activity is missing :

Based on
- Their partial activity pattern
- Their similarity with others in terms of connections and activity times

RECOVERING MISSING DATA FOR THE SPREADING PROCESS

▹ ***Factorization*** : extraction of mesoscale structures

- with structural composition [which links are involved]
- & temporal information [when it is active]

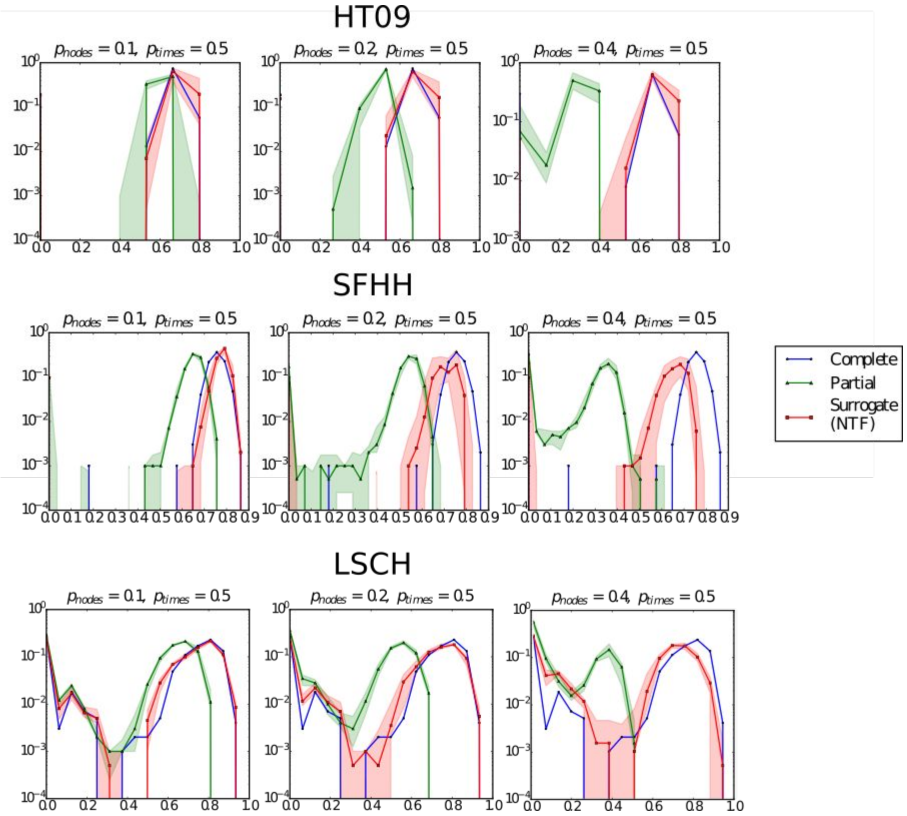
$\Longrightarrow$ approximated network with correct node activities

▹ "***Heterogenization*** "

- Correction of the weights according to the global distribution

$\Longrightarrow$ approximated network with heterogeneity properties (burstiness…)

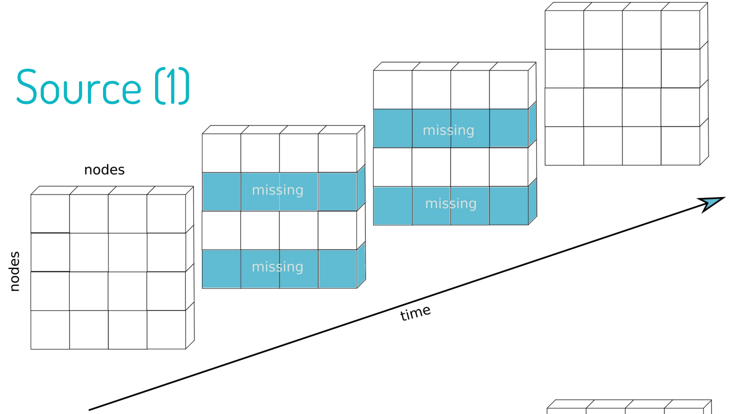
RESULTS : EPIDEMIC SIZE DISTRIBUTION



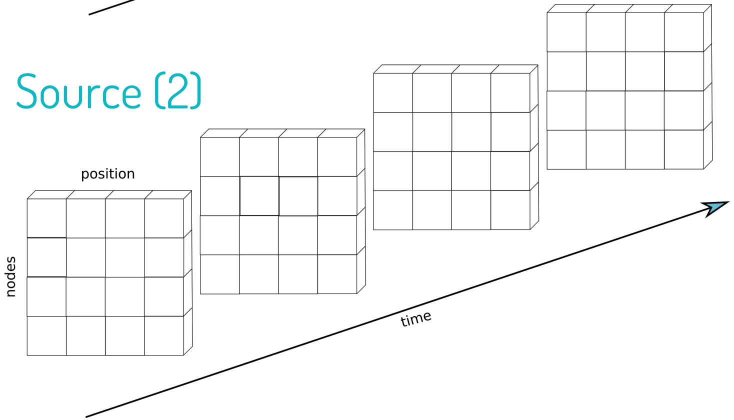
Sapienza, et al., Estimating the outcome of spreading processes on networks with incomplete information: a mesoscale approach, 2017.

RECOVERING MISSING DATA (2)

Source (1)



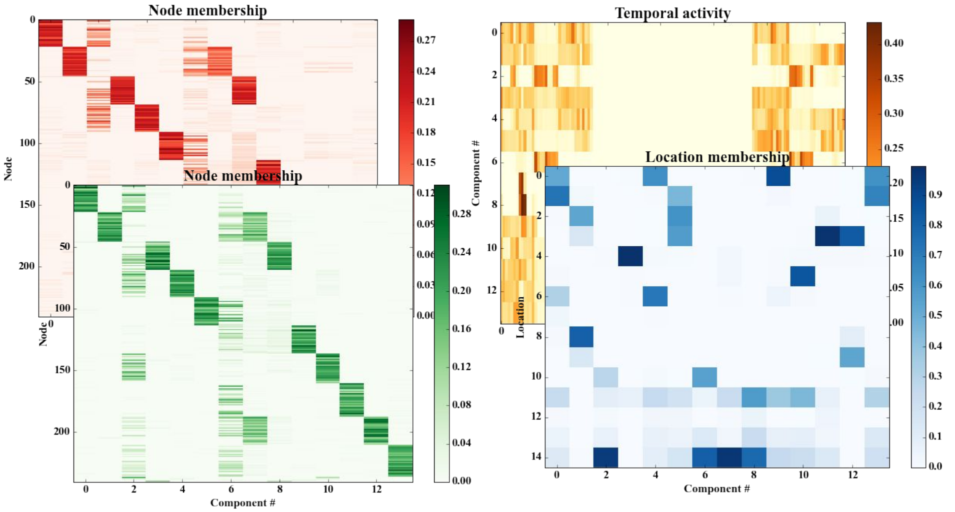
Source (2)



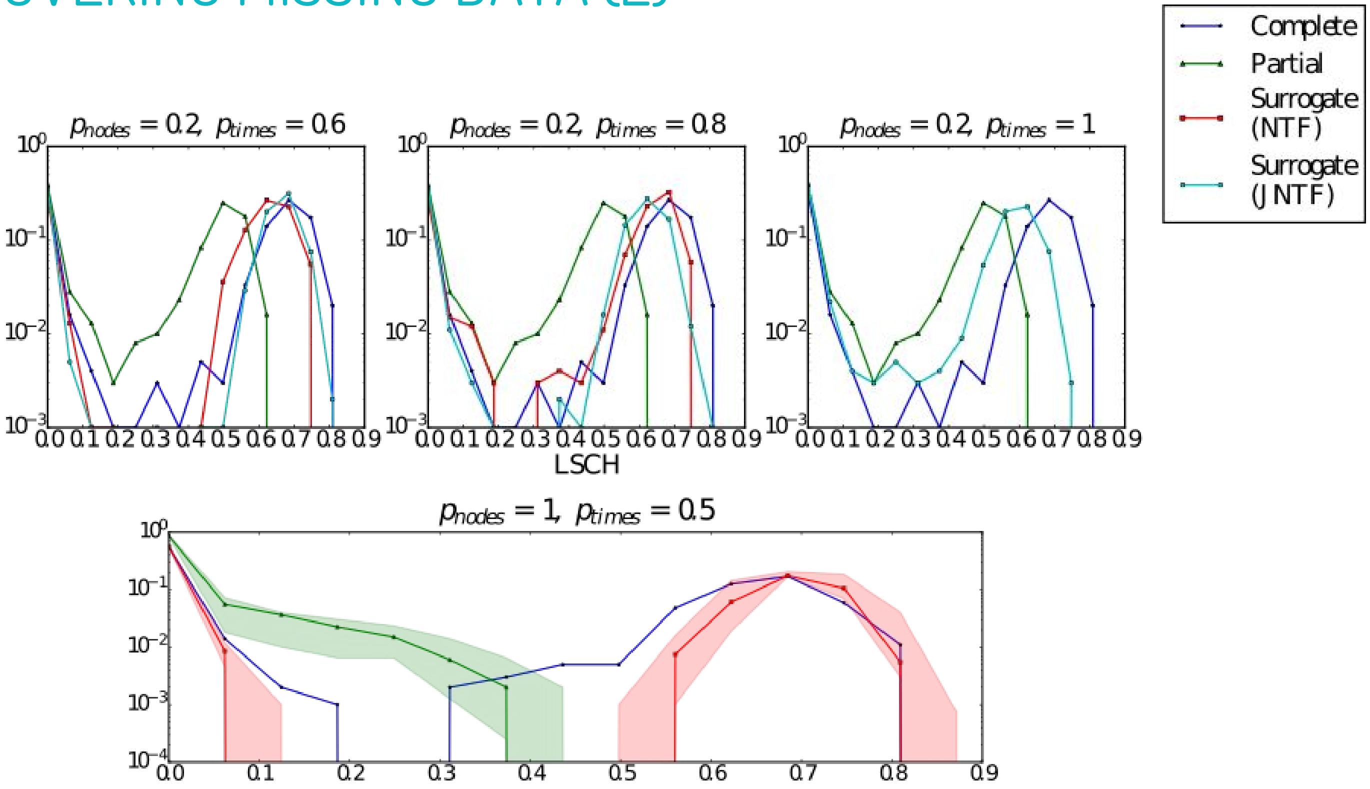
Joint-factorization of multiple sources

"Infer" activity of the nodes for which part of the activity is missing :

- $$\min \frac{1}{2} \sum_{k=1}^{K} \|\mathcal{T}_k - [\![\lambda_k, \mathbf{A}_k, \mathbf{B}_k, \mathbf{C}_k]\!]\|_F^2$$



RECOVERING MISSING DATA (2)



CONCLUSIONS ON RECOVERING DATA

We propose a technique to recover missing data based

- on factorization that efficiently recovers node activity

We adapted it by taking into account the need for heterogeneous distributions

- to recover the result of spreading processes [evolution and epidemic sizes]

We generalized to be able to merge the information from several data sources

No metadata were used

CONCLUSIONS

- Non-negative tensor factorization able to transform a network into an additive representation of meaningful structures

- Possible to handle missing values

- Framework easily extendable to multiple data sources

REFERENCES

- Detecting Anomalies in Time-Varying Networks Using Tensor Decomposition
  A Sapienza, J Wu, L Gauvin, C Cattuto
  2015 IEEE International Conference on Data Mining Workshop (ICDMW), 516-523
- Revealing latent factors of temporal networks for mesoscale intervention in epidemic spread
  L Gauvin, A Panisson, A Barrat, C Cattuto
  2015 arXiv preprint arXiv:1501.02758
- Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach
  L Gauvin, A Panisson, C Cattuto
  2015 PloS one 9 (1)
- Estimating the outcome of spreading processes on networks with incomplete information: a mesoscale approach
  A Sapienza, A Barrat, C Cattuto, L Gauvin
  2017

THANK YOU!

ALAIN BARRAT

CIRO CATTUTO

ANDRE PANISSON

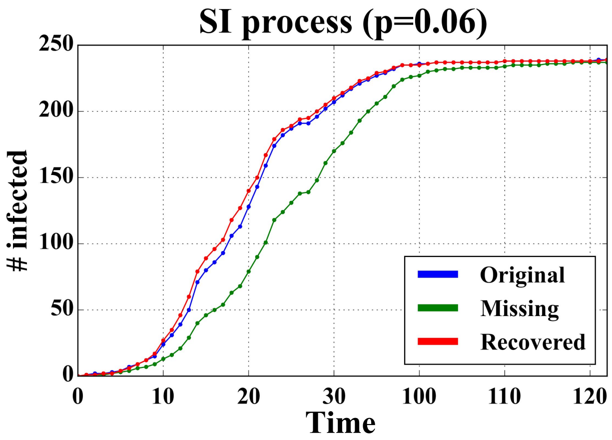
ANNA SAPIENZA

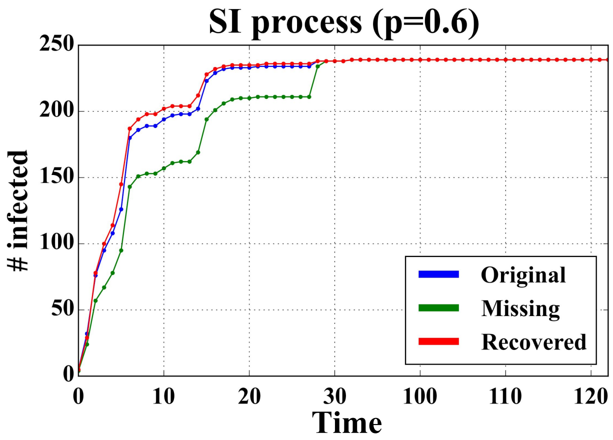
https://laetitiagauvin.github.io/

INSTITUTE
FOR SCIENTIFIC INTERCHANGE
FOUNDATION

RESULTS : SPREADING PROCESS EVOLUTION

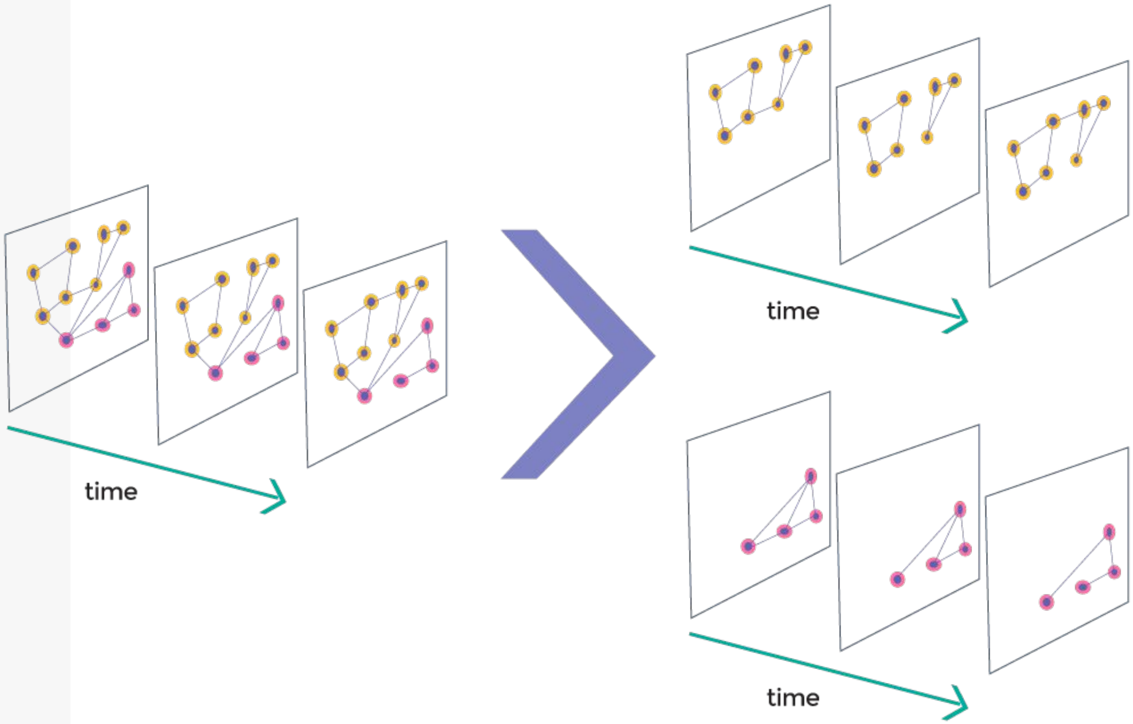
- 10% of nodes / 50% activity deleted
- 2 data sources : contacts + positions
- Joint-factorization + weight correction
- Susceptible-Infected on the approximated network





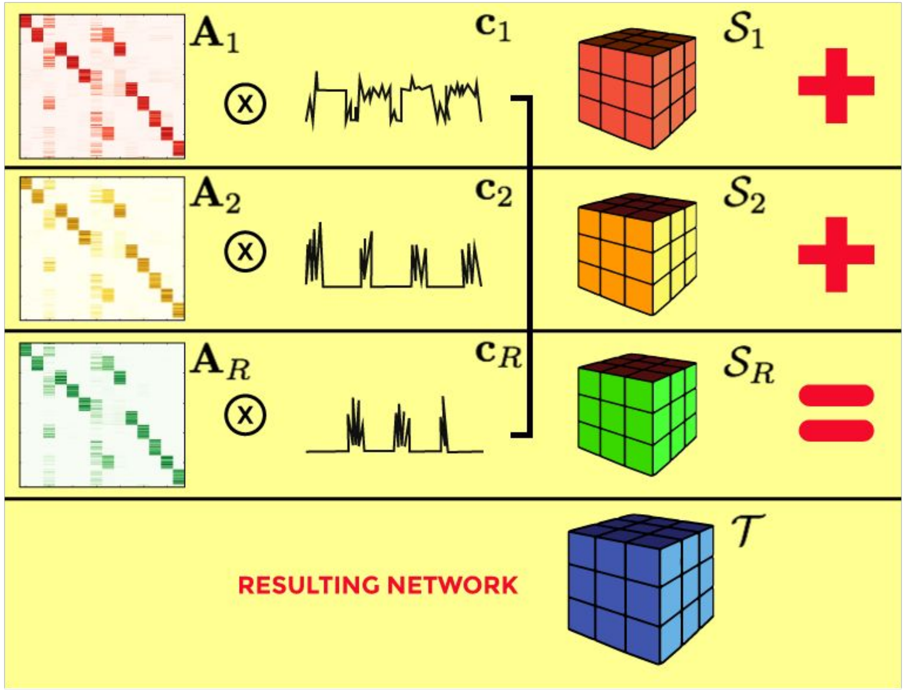
3.
GENERATIVE MODELS OF TEMPORAL NETWORKS

IDEA



A temporal network is built from sub-networks whose links have a correlated activity

MODEL

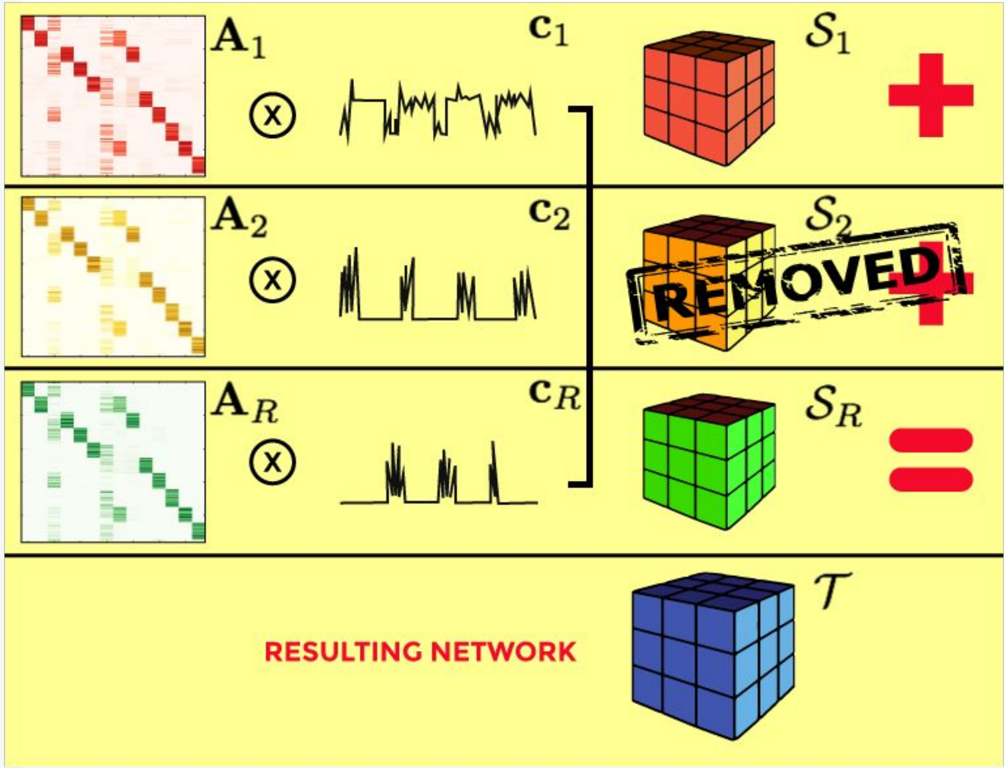


create a generative model where we can control separately temporal and topological structures and combine them

$$\mathcal{T} = \sum_{r=1}^{R} \mathbf{A}_r \circ \mathbf{c}_r$$

$$\mathbf{A}_r = \mathbf{a}_r \circ \mathbf{b}_r$$

IMPACT STUDY

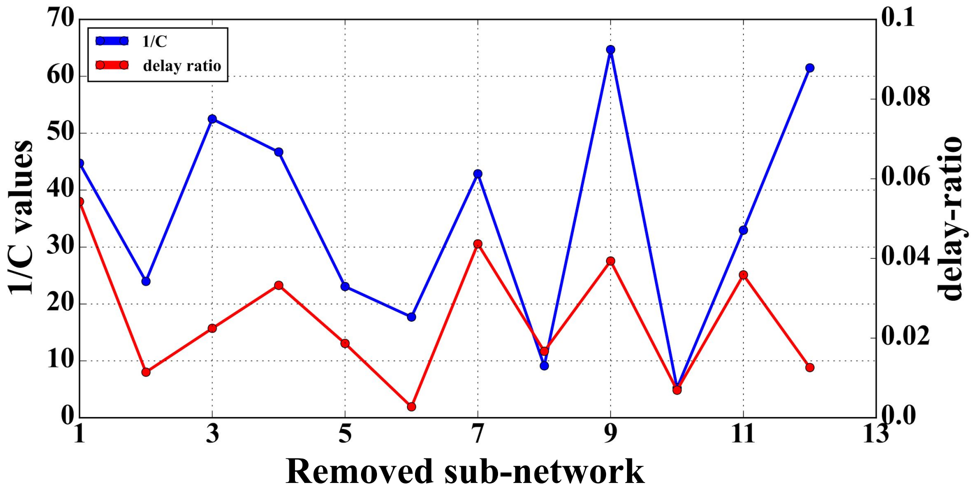


1. Generate a synthetic network
2. remove sub-networks one at a time
3. simulate an SI process over the original network and over the one given by the removal
4. compute the delay-ratio

IMPACT MEASURE

The impact of each sub-network is studied by the comparison between the delay ratio and the clustering coefficient:

$$C_i = \frac{\sum_{j,k} w_{ij} w_{jk} w_{ik}}{\sum_{j \neq k} w_{ij} w_{ik}}$$

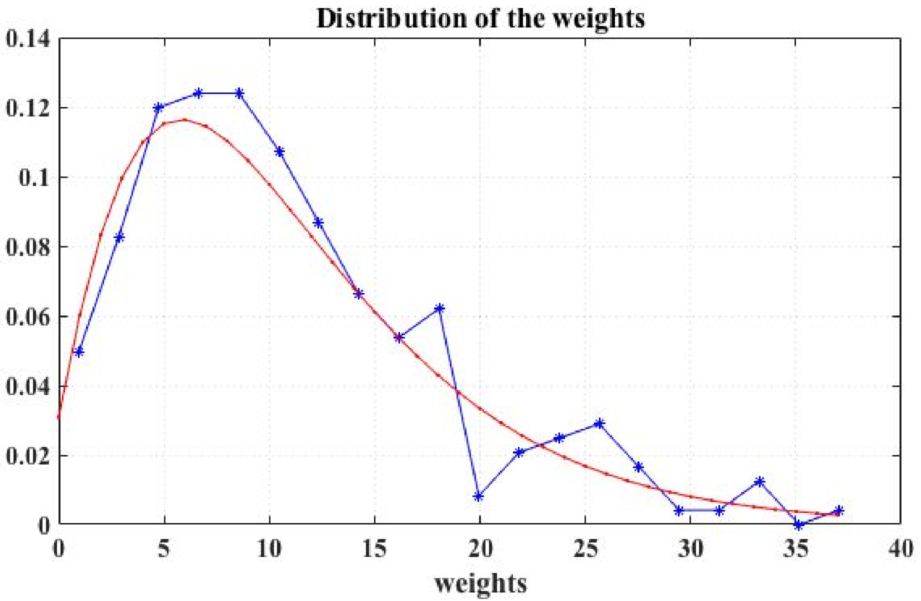


NEXT STEP

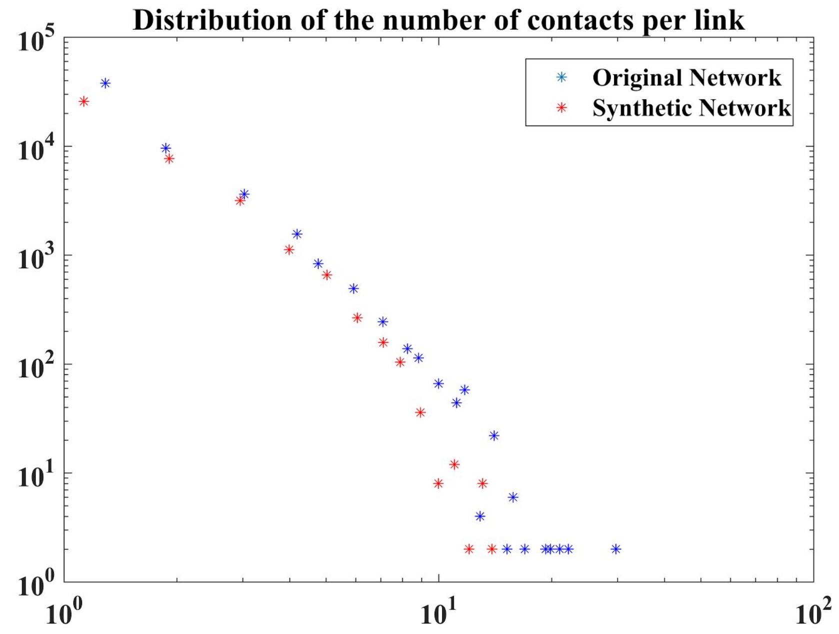
Use the negative binomial distribution

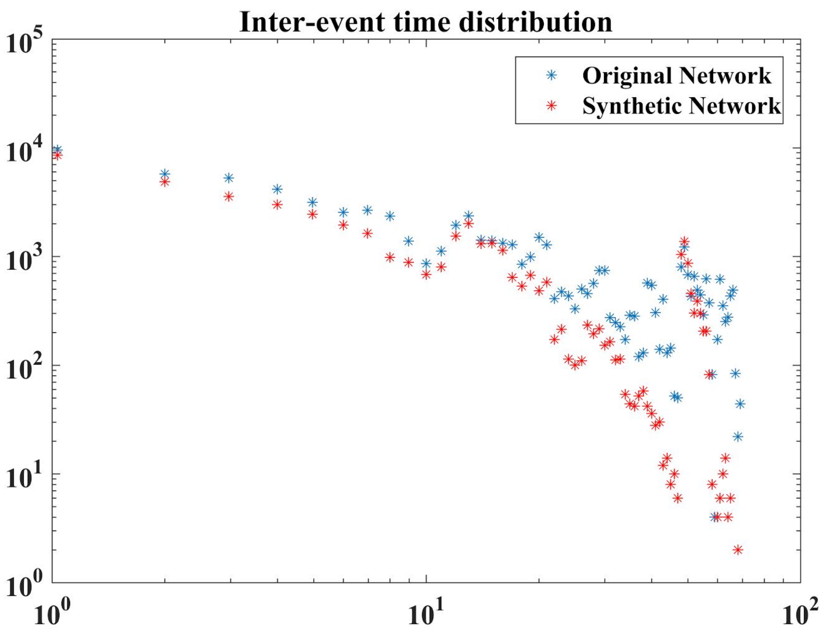
$$D\left(x\right)=\sum_{n=0}^{x}\binom{n+r-1}{r-1}p^{r}\left(1-p\right)^{n}$$

- to generate the structural part
- to make the temporal activity bursty



NEXT STEP


Distribution of the number of contacts per link
Original Network
Synthetic Network
10^5
10^4
10^3
10^2
10^1
10^0
10^0
10^1
10^2


Inter-event time distribution
Original Network
Synthetic Network
10^5
10^4
10^3
10^2
10^1
10^0
10^0
10^1
10^2

CONCLUSIONS

- Approach to the problem of studying the interplay between temporal network properties and dynamical processes
- Create temporal network in which we can control separately the temporal and topological properties
- Identification of the clustering coefficient value as a decisive factor to predict the impact on the process
- Next steps to make the model in a more principled way

RECOVERING MISSING DATA (1)

We solve a minimization problem to reveal mesoscale structures:

- Either we rebuild the tensor
- Or we just keep the following information:

  which links are involved in which structure (sub-network),

  when each structure is active

RESULTS : NODE ACTIVITIES

- 10% of nodes /50% activity deleted
- One data source : contacts
- Non-negative factorization on the tensor with missing values

Pearson coeff. *span*	[0.65,0.93]
median	0.84
p-value	$<10^{-3}$