Advanced Laboratory of Complex Network Analysis

Why? Learn to apply SNA in “research-like” contexts

Where? MS in Data Science

Who? Giulio Rossetti & Barbara Guidi

Teachers

Prof. Giulio Rossetti
giulio.rossetti@isti.cnr.it

When: upon appointment
Where: Online

Prof. Barbara Guidi
barbara.guidi@unipi.it

When: upon appointment
Where: Online

Tutors

Each tutor will propose, follow, coordinate and support one (or more) group projects.

Data will be provided by the tutors as well.

Course Materials

E-learning:

- Lessons schedule
- Slides
- Announcements
- https://elearning.di.unipi.it/

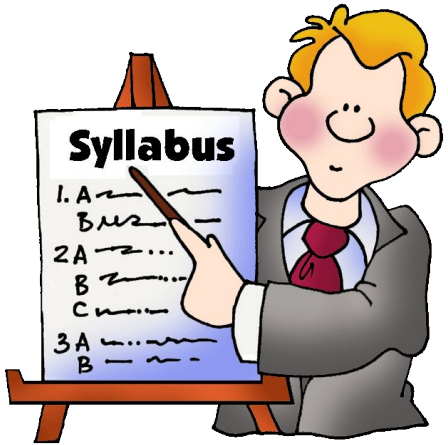
GitHub Repositories:

- Final Project
- Tutorials
- https://github.com/sna-unipi

Books:

- Coscia:
  The Atlas for Aspiring Network Scientists
- Barabasi
  Network Science
- Hiebert et al.
  Doing Research: A New Researcher's Guide

General Outline



Lectures (18h):

Formulating Hypotheses
From Simple Graphs to Advanced Models
Data Collection: Network Sampling
Data Collection: API & Web Scraping
Graph Transformation: Backboning
Enriching Topologies: Feature-rich modeling
How to Validate: statistical significance of network based analyses
Experiments from A to Z: Case studies

Laboratory (18h):

Each lecture will be complemented by a practical lab session
to familiarize with specific tasks and network analysis tools/libraries.

Group Project (10h):

Final Project assigned at the beginning of the course
Each group will be supervised by a tutor
Group-dedicated laboratory sessions

Exam & FAQs

Group Project + Oral Discussion

- Ideally perfected during the course
- Discussion upon appointment

Projects Presentation

- Public presentation to the project to peers

Is it required to have verbalized the SNA exam to join the course?
Not at all!
You can join even if you're still working on the SNA project
(and use the labs to get support for it in the meantime)

Is it mandatory to have followed the SNA course?
No, but it is (strongly) suggested.
SNA applications (e.g., CD, LP… will not be reintroduced)

Chapter 1

Recap: where were we?

Summary

- Networks, networks everywhere!
- Characterization
- Applications

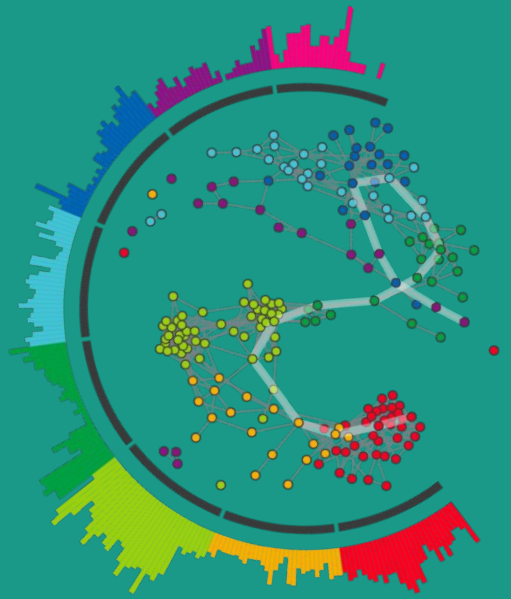
Reading

- “Network Science”, Barabasi

In the last season of SNA...

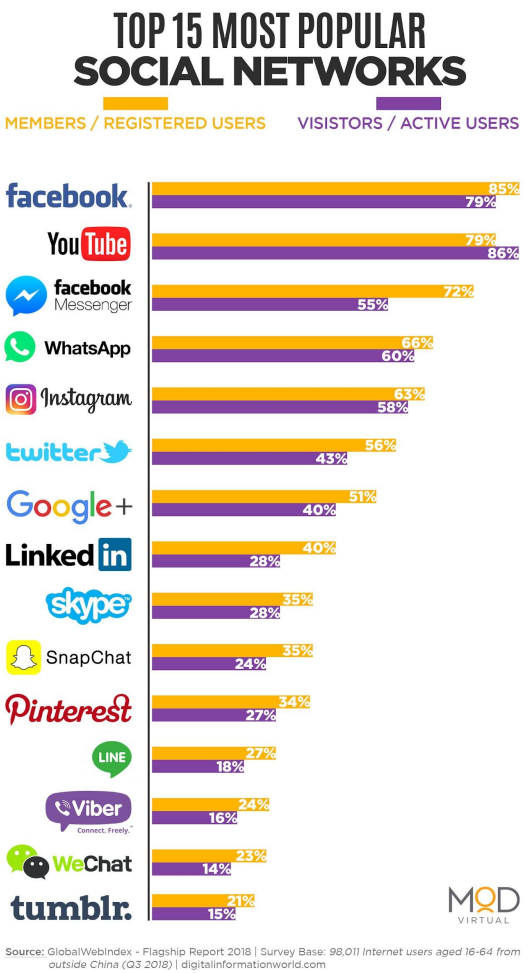
The role of networks

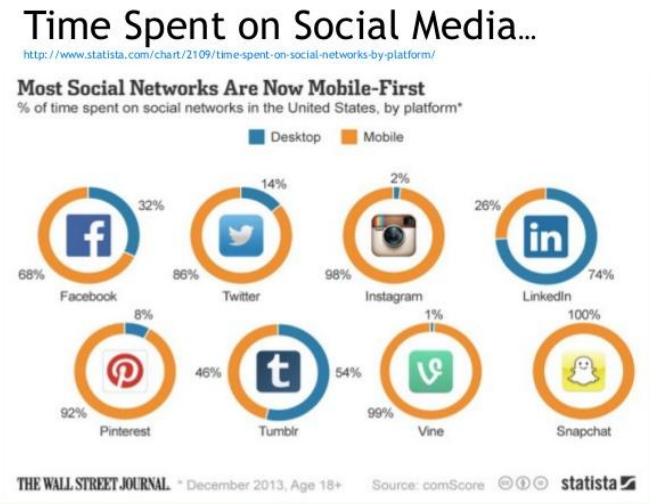
Behind each system studied in complexity there is an intricate wiring diagram, or a **network**, that defines the interactions between the component.

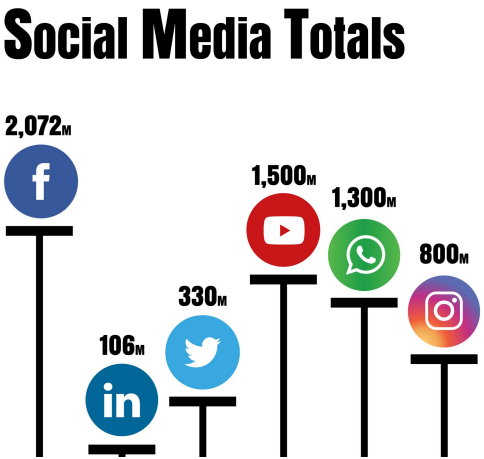


We will never understand **complex system** unless we map out and understand the networks behind them.

(Online) Social Networks







Central quantities in Network Science

Degree Distribution

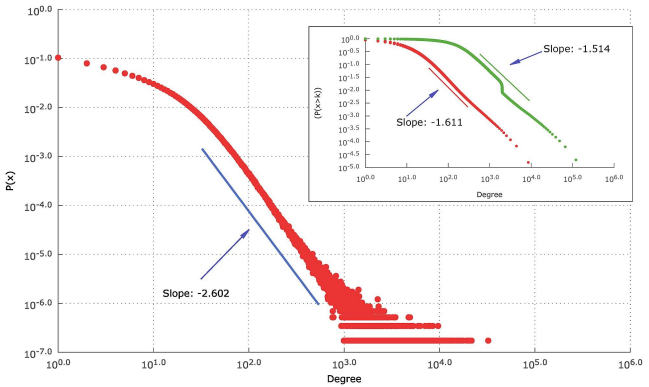
$P(k)$

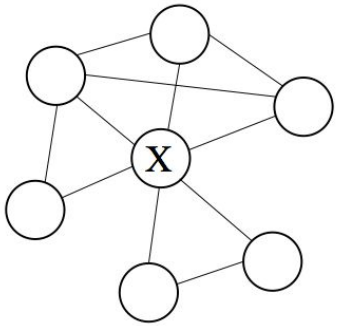
Path length

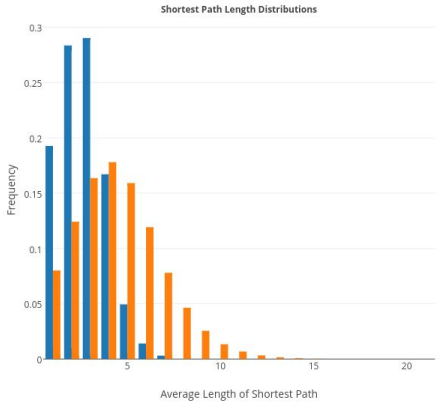
$\langle d \rangle$

Clustering Coefficient

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$







We learnt to characterize networks...

Network models
in a Nutshell

"All models are wrong, but some are useful"

- ER models and Configuration models are used as reference models in a very large number of applications
- WS, BA are more *"making a point"* type models: simple processes can explain some non-trivial properties of networks, unfound in random networks.
- Correlation is not causation.
  Are these simple processes the "cause"?
  Maybe, maybe not, sometimes...

Network	Degree Distribution	Path Length	Clustering Coefficient
Real-world networks	Broad	Short	Large
ER graphs	Poissonian	Short	Small
Configuration model	Custom, can be broad	Short	Small
Watts & Strogatz (in SW regime)	Poissonian	Short	Large
Barabasi Albert (Scale-Free)	Power-Law	Short	Rather Small
Other models	Power-law	Short	Large

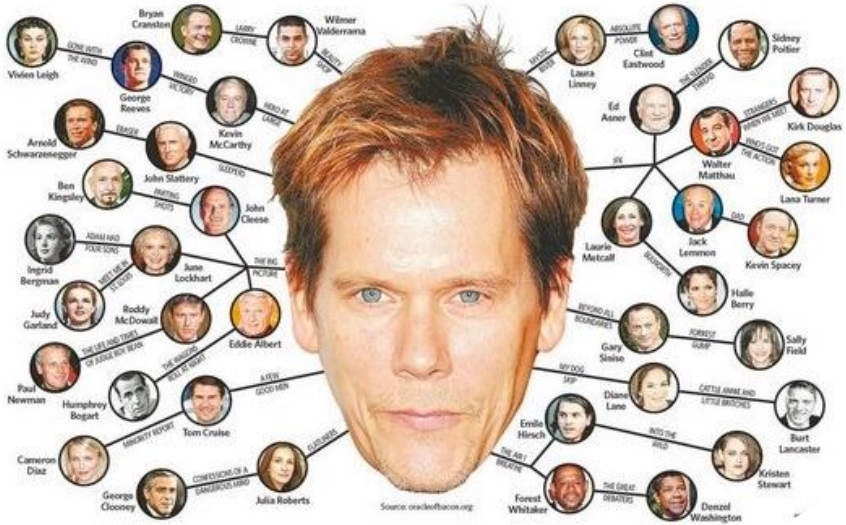
How important is a node in a network?

We can measure nodes importance using so-called centrality.

Bad term:
nothing to do with being central in general

Usage:

- Some centralities have straightforward interpretation
- Centralities can be used as node features for machine learning on graph



https://oracleofbacon.org/

Each centrality measures is a proxy of an underlying network process.

If such a process is irrelevant for the network than the centrality measure makes no sense

- E.g. If information does not spread through shortest paths, betweenness centrality is irrelevant

Centrality measures should be used with caution (a) for exploratory purposes and (b) for characterisation

Homophily

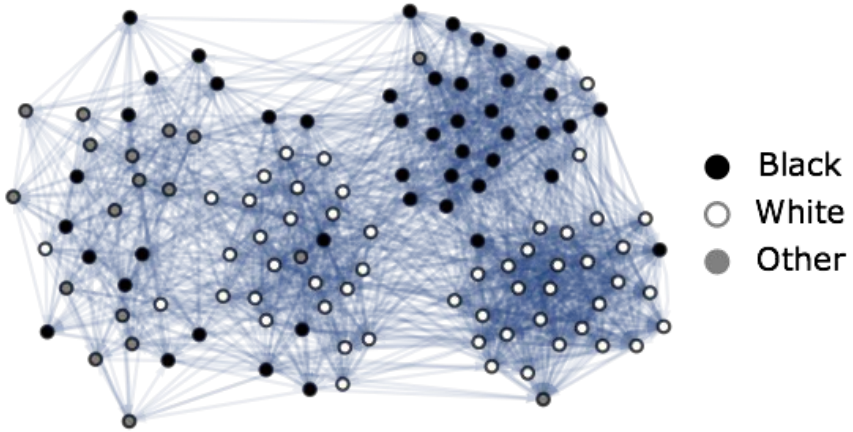
Property of (social) networks that nodes of the same attitude tends to be connected with

a higher probability than expected

- It appears as correlation between vertex properties of x(i) and x(j) if (i,j) ∈ E

Disassortative mixing:

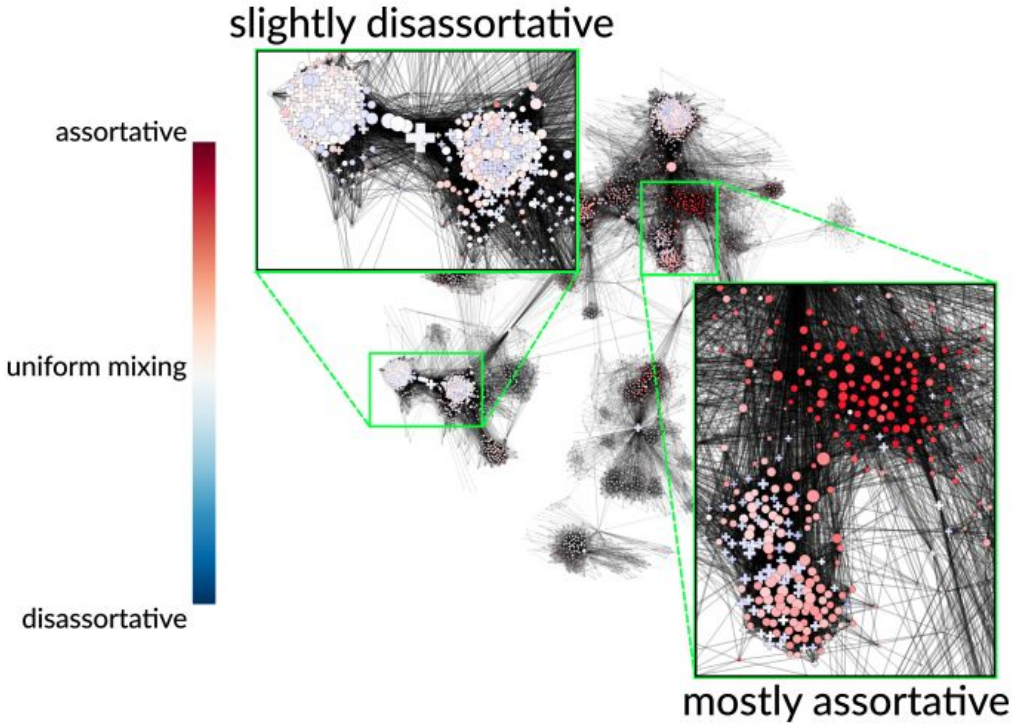
Contrary of homophily: dissimilar nodes tend to be connected
(e.g., sexual networks, predator-prey)



Examples of Vertex properties

age, gender, nationality,
political beliefs, socioeconomic status,
obesity, ...

Homophily can be a link creation mechanism or consequence of social influence (and it is difficult to distinguish)



Local assortativity of gender in a sample of Facebook friendships (McAuley and Leskovec 2012).

Different regions of the graph exhibit strikingly different patterns, suggesting that a single variable, e.g. global assortativity (Newman's), would provide a poor description of the system.

Limits of a global assortativity score

We discovered that not all edges are equals...

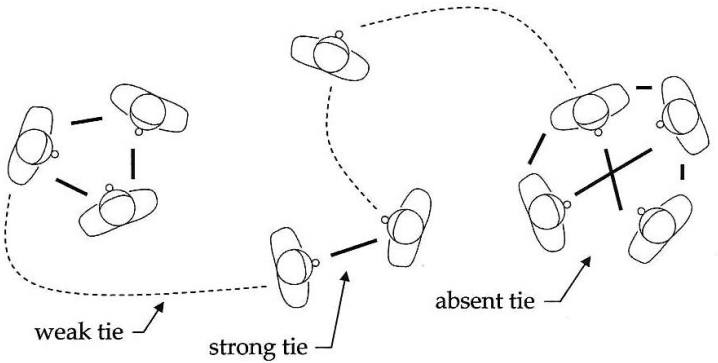
The strength of weak ties

Mark S. **Granovetter**, 1973

- (PhD Thesis)
  "How people get to know about new jobs?"
- Answer: Through personal contacts

Unexpected result:

Often acquaintances, **not** close friends… but *why*?



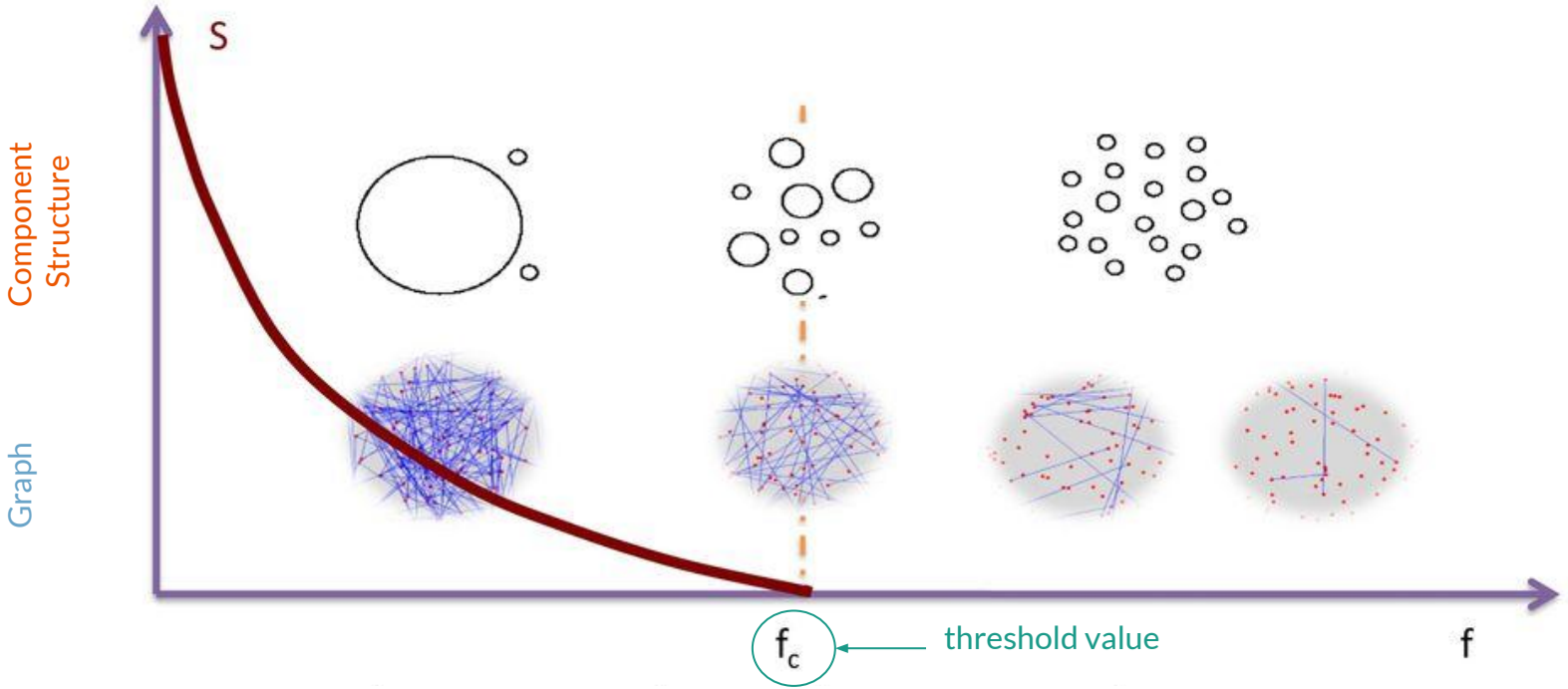
How to measure tie strength?

Granovetter's **dimensions of tie strength**:

- the amount of time spent interacting with someone,
- the level of intimacy,
- the level of emotional intensity,
- and the level of reciprocity.

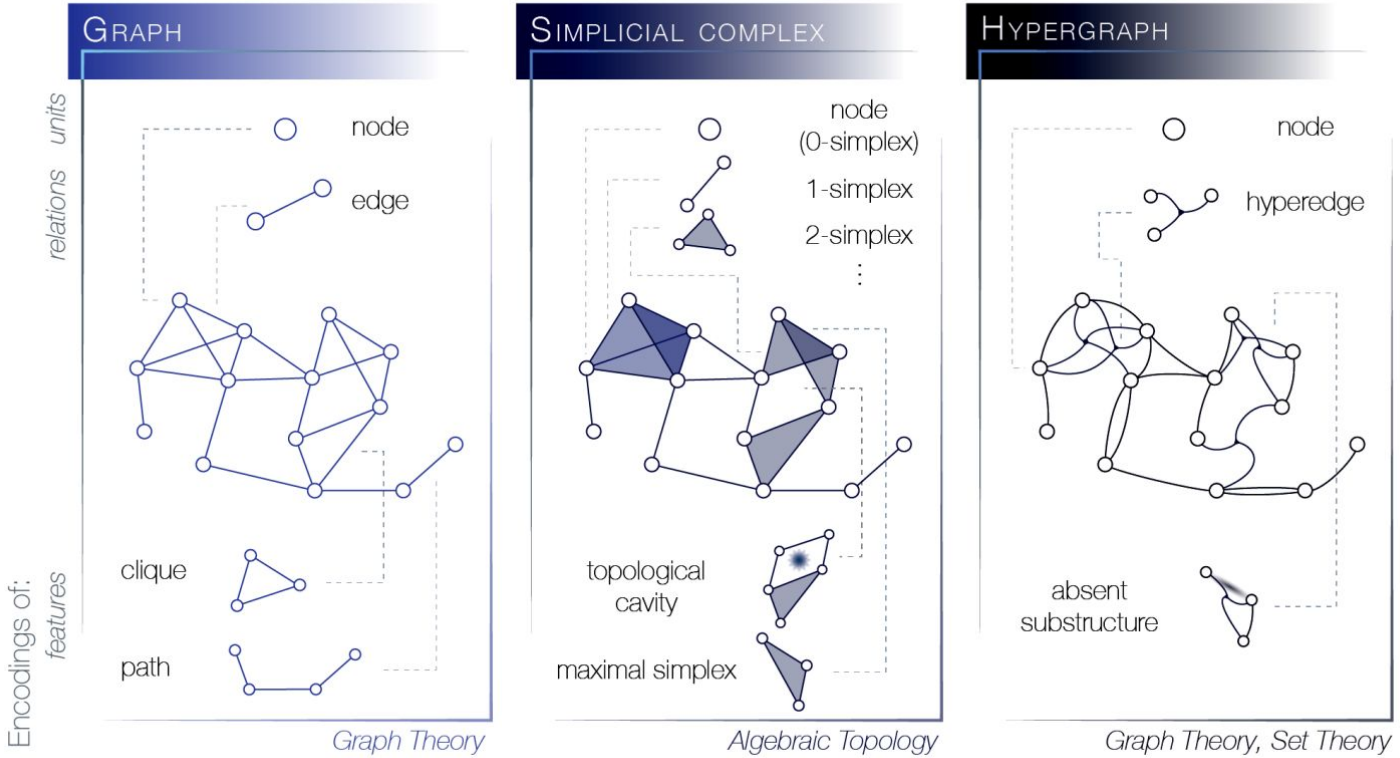
Granovetter, Mark S. "The strength of weak ties." Social networks. Academic Press, 1977. 347-367.

f = fraction of removed nodes

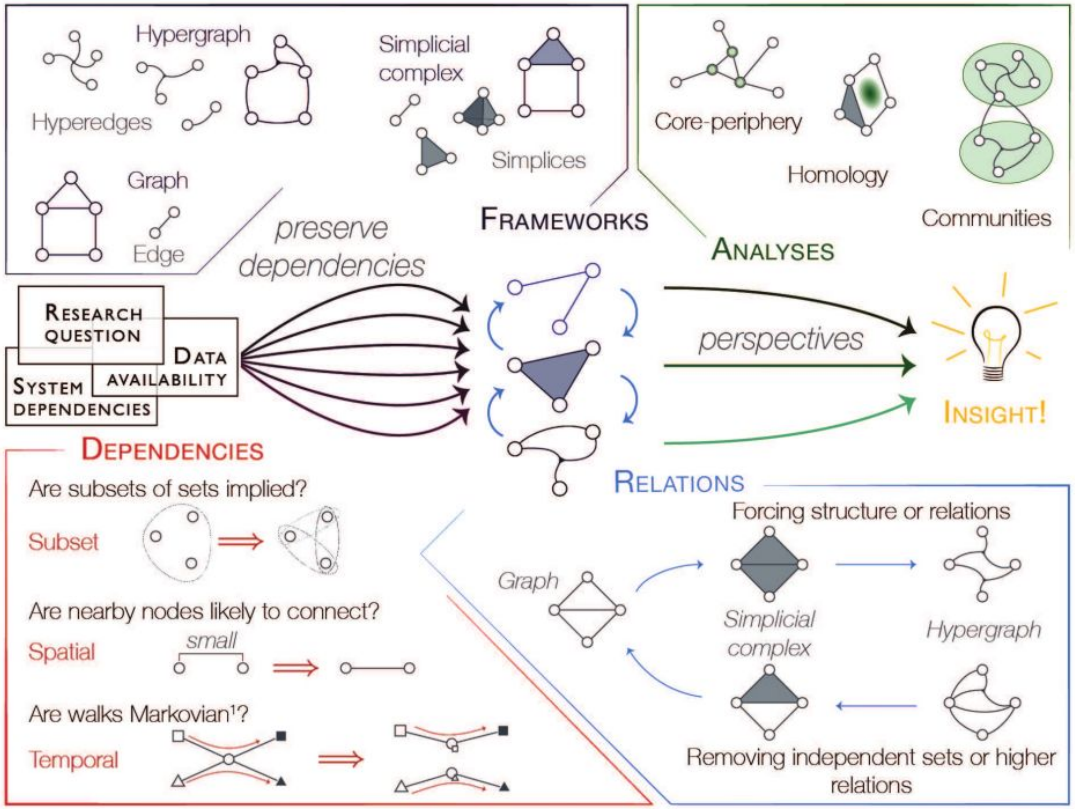


Topology affects resilience

… and that edges are often not enough…



Different complex systems require different modeling frameworks…



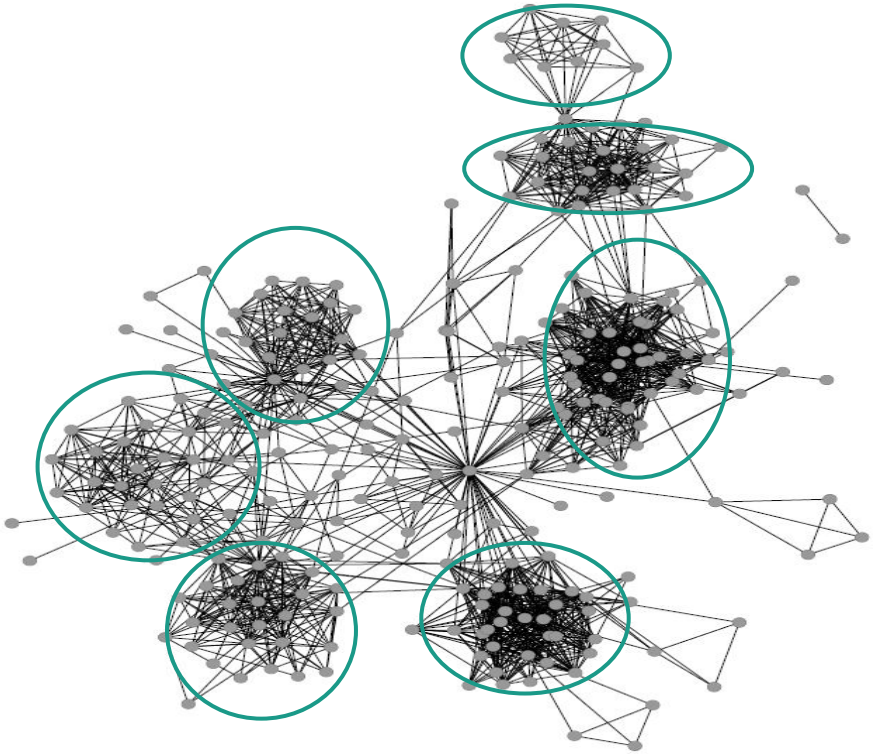
… explicitating peculiar entities' dependencies and enabling specific analysis!

We divided to conquer...

Communities in Complex Networks

In simple, small, networks it is easy identify them by looking at the structure…

- i.e., using a Force directed layout



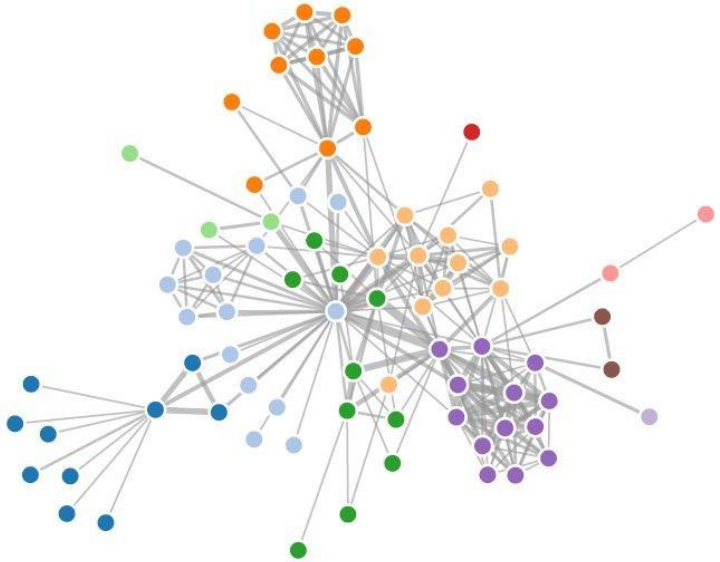
Community Discovery is, perhaps, the hottest topic in complex network analysis

Major issues:

- Problem definition
- Community evaluation

Problem specializations:

- Evolutionary Community Discovery
  (How do communities evolve in dynamic networks?)
- Multidimensional Community Discovery
- ...



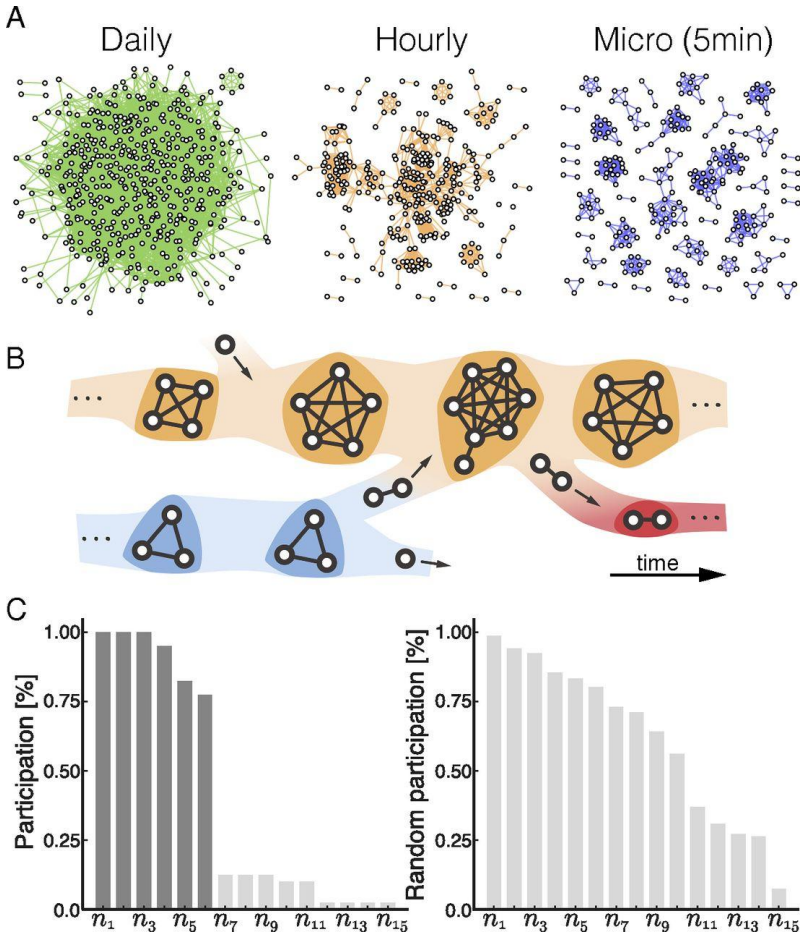
… and start questioning about time!

Why bother of time?

Most real world networks are dynamic

- Facebook friendship
  - People joining/leaving
  - Friend/Unfriend

- Twitter mention network
  - Each mention has a timestamp
  - Aggregated every day/month/year => still dynamic



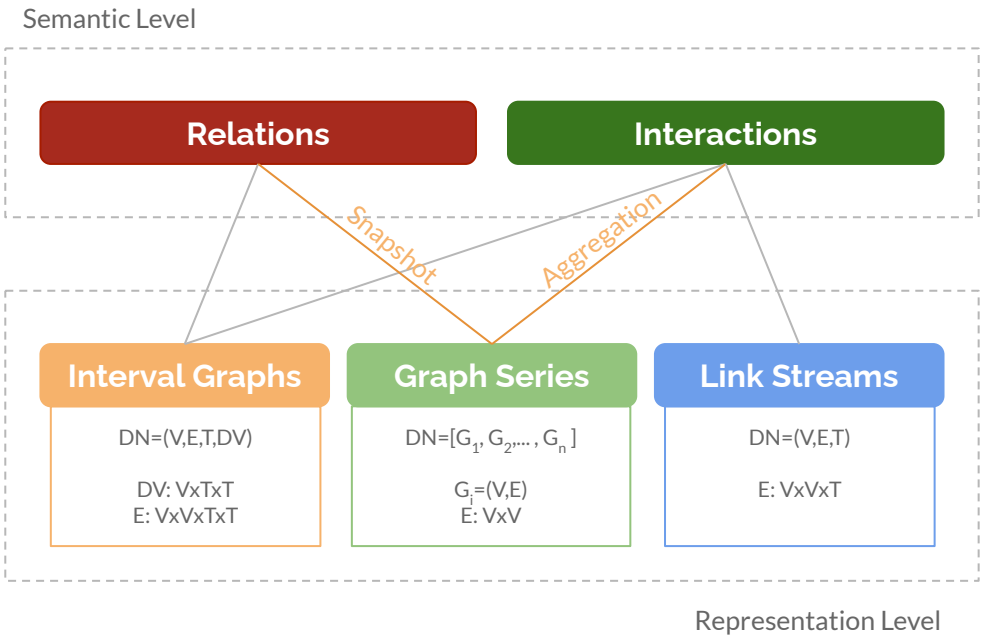
Semantics and how to represent them

Relations

The graph is more and more stable, until most observations are completely similar to previous/later ones (frequency faster than change rate)

Interactions

The graph is less and less stable, until each observation is a graph in itself, thus completely different from previous/later ones (frequency faster than observed events rate)



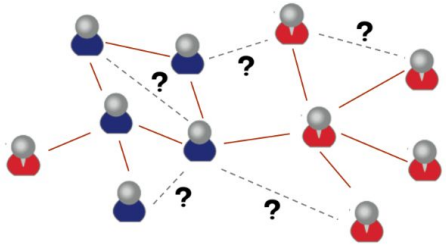
Link Prediction

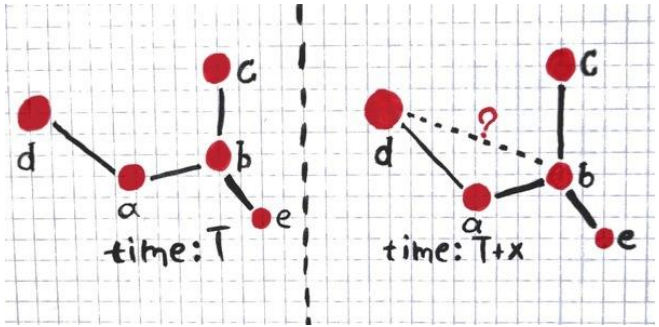
Goal

Understanding how networks evolve

Problem definition

Given a snapshot of a network at time **t**, (accurately) predict the edges that will appear in the network during the interval **(t, t+1)**





Liben-Nowell, David, and Jon Kleinberg. "The link-prediction problem for social networks." Journal of the American society for information science and technology 58.7 (2007): 1019-1031.

Communities
In Dynamic Networks

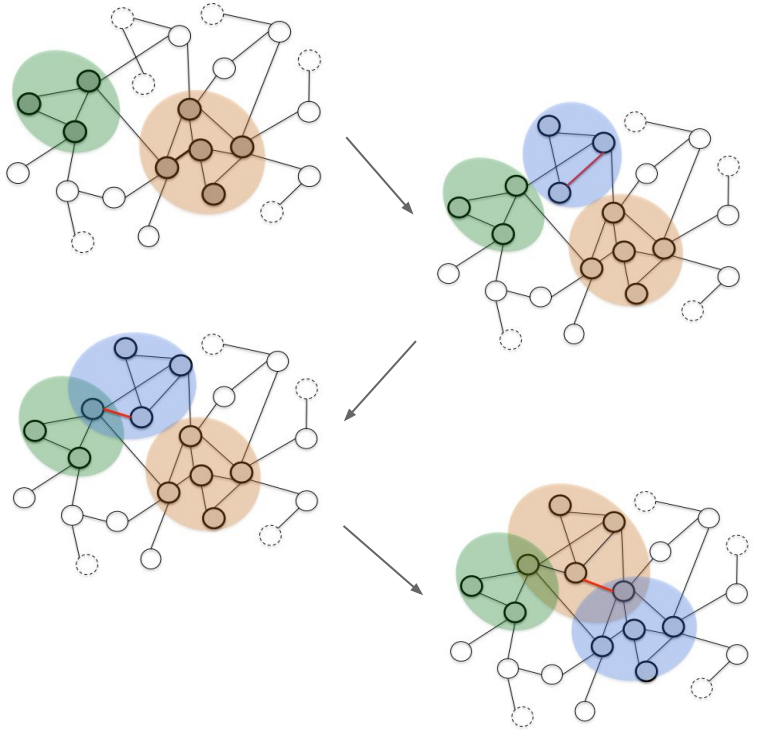
Networks change with time…

- Nodes appear and vanish
- Edges appear and vanish

…communities must change too!

DCD:

identify/track changes in community structure



Cazabet, Remy, and Giulio Rossetti. "Challenges in community discovery on temporal networks." Temporal Network Theory. Springer, Cham, 2019. 181-197.

A Novel Problem:

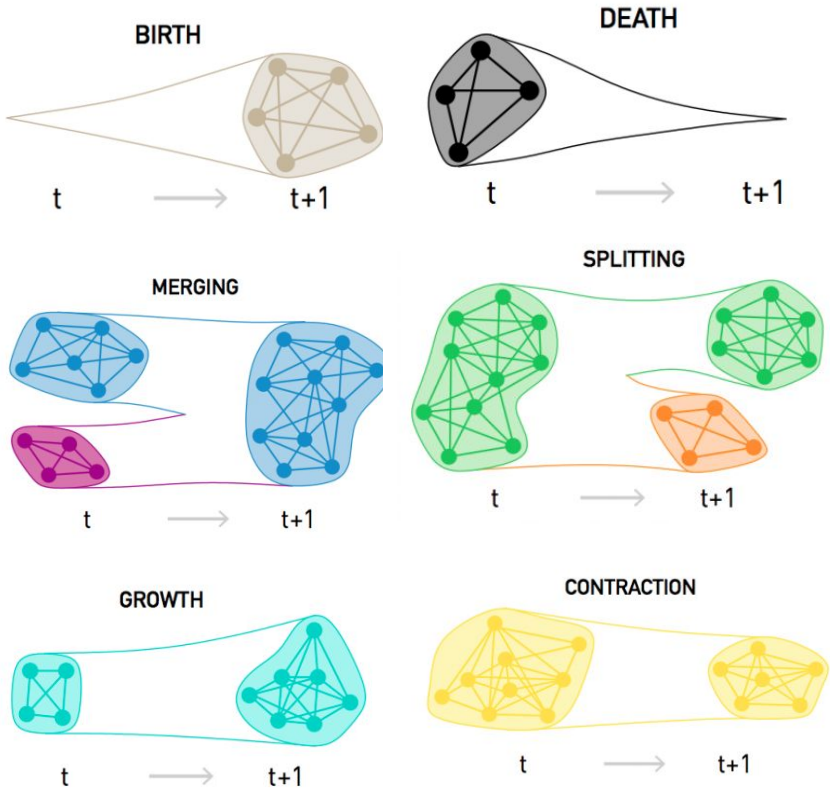
Community life-cycle tracking

As time goes by the rising of novel nodes and edges (as well as the vanishing of old ones) led to network perturbations

Communities can be deeply affected by such changes

Three main strategies:

- Identify & Match
- Informed Iterative algorithms
- Stable Identification



Finally we learnt to model dynamic phenomena

Network effects

Diffusion happens only when the carries of the diseases/virus/idea are connected to susceptible nodes.

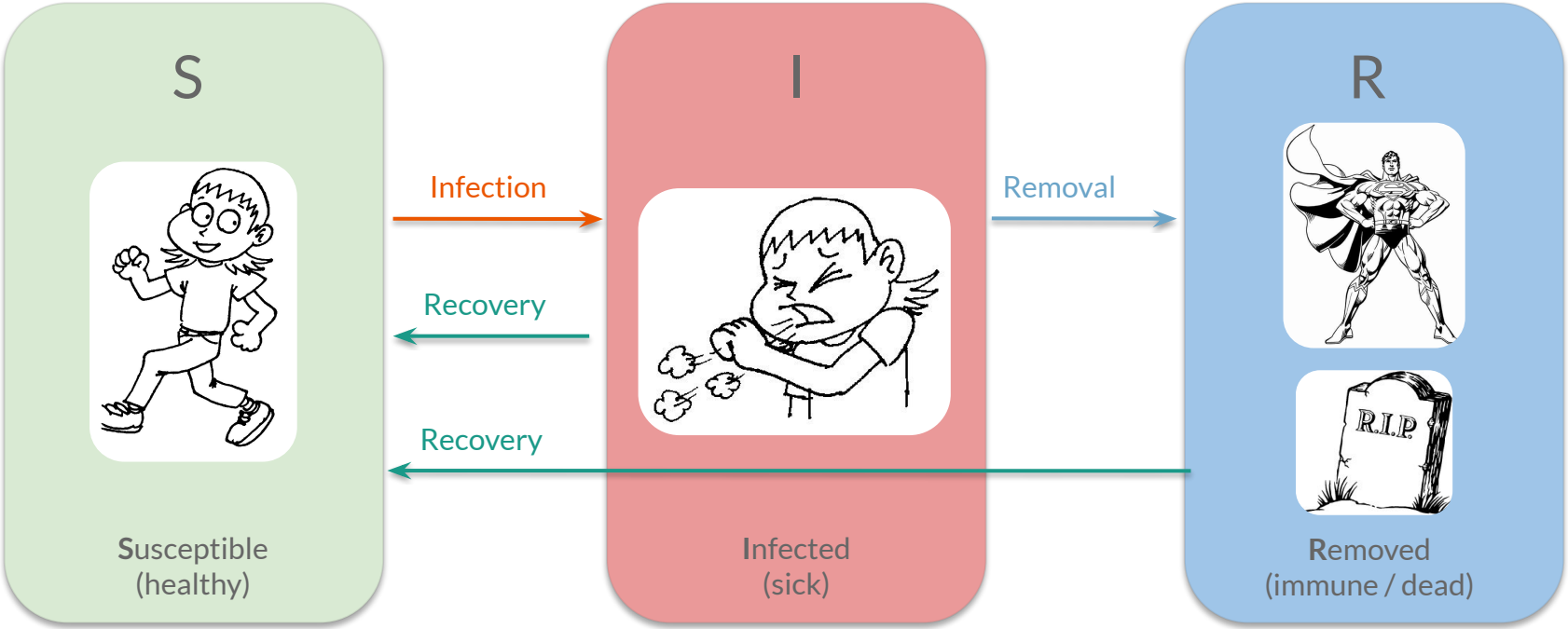
Diffusive phenomena can modeled describing:

- “node statuses”
- “transition rules”

We will always start from a MF model, then extend it to a networked context.

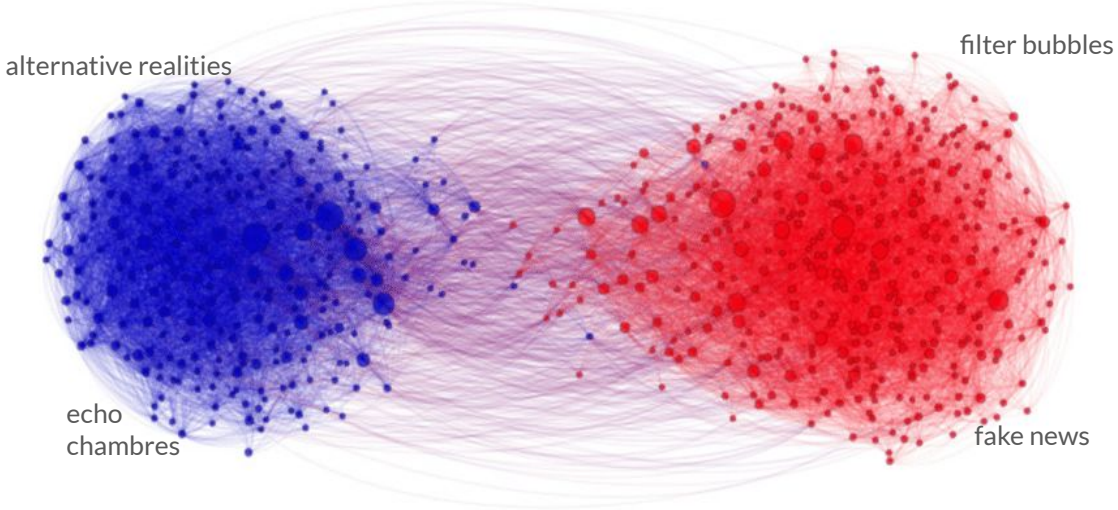
Spreading speed & patterns will vary depending on:

- model’s parameter values (as in mean field)
- Initial infection seeds
- Topology that surrounds infected nodes (e.g., communities may act as barriers)
- Degree of homophily of connected nodes
- ...



Epidemic Modeling

Polarization of the public debate



Adamic, Lada A., and Natalie Glance. "The political blogosphere and the 2004 US election: divided they blog." ACM (2005).

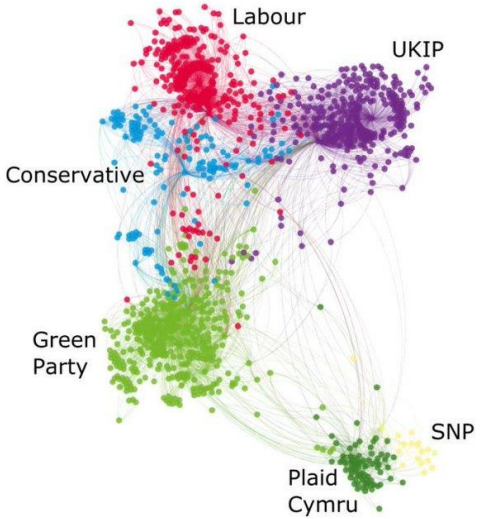
Algorithmic Bias

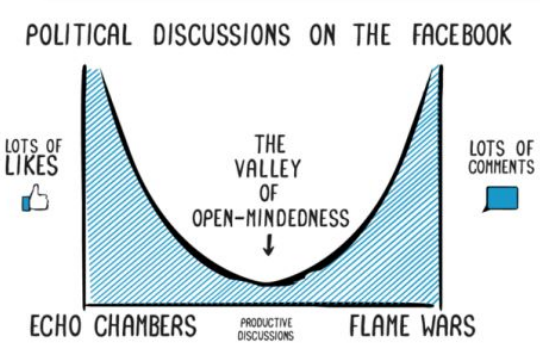
Is this the whole story?

Unfortunately, it is not.

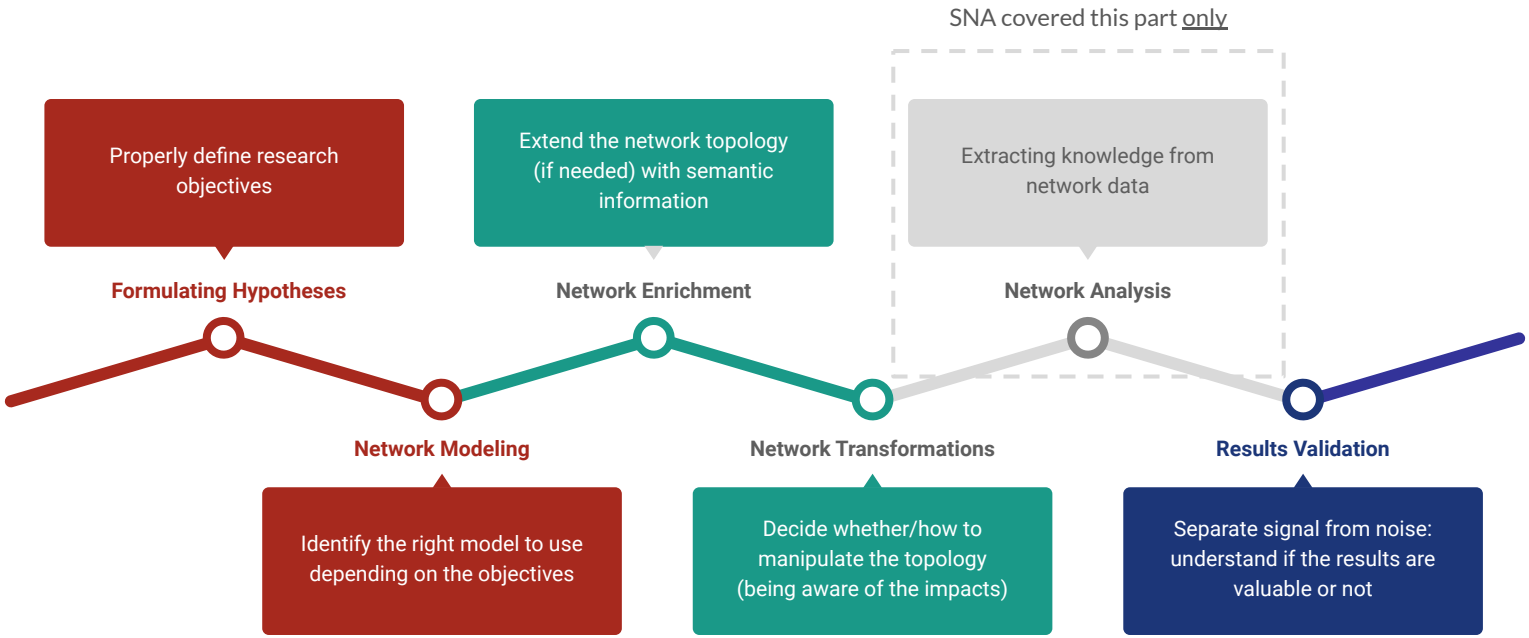
The situation in reality is even worse

- Simulations performed in mean field
- The observed effects can be exacerbated by the topology of the social network





… and now, what?


SNA covered this part only
Properly define research objectives
Formulating Hypotheses
Identify the right model to use depending on the objectives
Network Modeling
Extend the network topology (if needed) with semantic information
Network Enrichment
Decide whether/how to manipulate the topology (being aware of the impacts)
Network Transformations
Extracting knowledge from network data
Network Analysis
Separate signal from noise: understand if the results are valuable or not
Results Validation

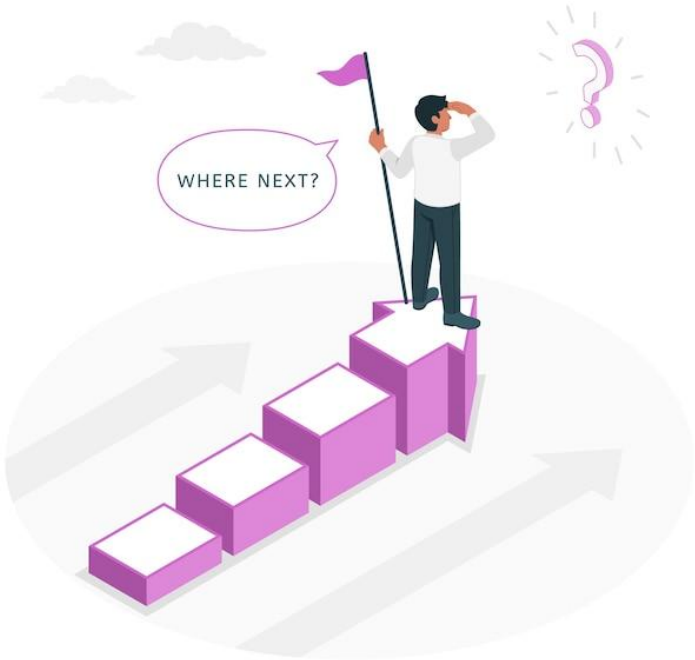
Practical stuff

Organize groups!

Each lecture will be followed by one (or more) laboratory focused on hands-on related tutorial exercises.

Next lecture we will discuss how to formulate proper hypotheses.
The following laboratory will be focused on:

- Introducing you the available “projects”
- Assign project (and tutors) to groups
- Work together to propose some preliminary hypotheses to be tested



Chapter 1

Conclusion

Take Away Messages

1. We know the basis, now let's move to some advanced stuff!

What's Next

Chapter 2: Formulating Hypotheses