Network Analysis and Mining
11. Spreading phenomena

Maximilien Danisch, Lionel Tabourier

LIP6 – CNRS and Sorbonne Université

first_name.last_name@lip6.fr

January 5 2021

Outline

About this section

Initially: from Part A (Complex Networks Analysis tools)
$\rightarrow$ to Part B (Graph mining)

- use our knowledge and tools to explore real networks
- adapt them to specific problems

$\rightarrow$ problem-oriented view

Illustration: a few Social Network Analysis concepts

Applying our tools to Social Network Analysis
Epidemic spreading models on graphs
The homophily phenomenon
Local density and community structure
Impact on social contagion

About this section

Initially: from Part A (Complex Networks Analysis tools)
$\rightarrow$ to Part B (Graph mining)

- use our knowledge and tools to explore real networks
- adapt them to specific problems

$\rightarrow$ problem-oriented view

Illustration: a few Social Network Analysis concepts

3/18

What is homophily?

From Greek, *homo:* same, similar and *philos:* friend of, to like
$\rightarrow$ *"birds of a feather flock together"*

Observed for a long time in sociology:

- smoking habits, food habits
- residential segregation
- voting behavior
- ...

Do we observe this phenomenon in Social Networks Analysis?

What is homophily?

From Greek, *homo:* same, similar and *philos:* friend of, to like
$\rightarrow$ *"birds of a feather flock together"*

Observed for a long time in sociology:

- smoking habits, food habits
- residential segregation
- voting behavior
- ...

Do we observe this phenomenon in Social Networks Analysis?

Structural homophily, based on degree

Degree-based assortativity

Do high-degree nodes connect to high-degree nodes?

Measured using:

- q_k probability distribution of the remaining degree:
  if p_k is the degree distribution then $q_k = \frac{(k+1)p_{k+1}}{\sum_j j.p_j}$
- e_{jk} joint probability of remaining degree distribution (RDD),
  i.e. probability to pick an edge which ends have remaining degree j and k

We measure the assortativity coefficient r:

$$r = \frac{\sum_{j,k} j.k.(e_{jk} - q_j q_k)}{\sigma_q^2} \text{ with } \sigma_q^2 = \frac{1}{n} \sum_k q_k (k - \overline{k})^2$$

Structural homophily, based on degree

Degree-based assortativity

Do high-degree nodes connect to high-degree nodes?

Measured using:

- q_k probability distribution of the remaining degree:
  if p_k is the degree distribution then $q_k = \frac{(k+1)p_{k+1}}{\sum_j j.p_j}$
- e_{jk} joint probability of remaining degree distribution (RDD),
  i.e. probability to pick an edge which ends have remaining degree j and k

We measure the assortativity coefficient r:

$$r = \frac{\sum_{j,k} j.k.(e_{jk} - q_j q_k)}{\sigma_q^2} \text{ with } \sigma_q^2 = \frac{1}{n} \sum_k q_k (k - \overline{k})^2$$

Structural homophily, based on degree

Degree-based assortativity

Do high-degree nodes connect to high-degree nodes?

We measure the assortativity coefficient r:

$$r = \frac{\sum_{j,k} j.k.(e_{jk} - q_j q_k)}{\sigma_q^2} \text{ with } \sigma_q^2 = \frac{1}{n}\sum_k q_k (k - \overline{k})^2$$

- r measures if nodes connected together in the graph have more similar degrees than expected (if connected randomly)
- σ_q^2 is the variance of the remaining degree distribution (RDD) q_k
- r is a normalized quantity ($\in [-1:1]$)
- r: ratio between covariance across links of RDD and variance of RDD

Newman - *Phys. Rev. E, 2003*

Structural homophily, based on degree

Typical degree-assortativity on social networks:

- astrophysics coauthorship: $r = 0.235$ (Georgia Tech data)
- actor collaboration: $r = 0.227$ (Notre-Dame Univ data)
- friendship network: $r = 0.039$ (Livejournal data)

$\rightarrow$ social networks are usually degree-assortative

Note that all complex networks are not degree-assortative:

- Internet AS level: $r = -0.215$ (UCLA data)
- human protein network: $r = -0.126$ (Vidal data)
- US power grid network: $r = 0.003$ (Tore Opsahl data)

Structural homophily, based on degree

Typical degree-assortativity on social networks:

- astrophysics coauthorship: $r = 0.235$ (Georgia Tech data)
- actor collaboration: $r = 0.227$ (Notre-Dame Univ data)
- friendship network: $r = 0.039$ (Livejournal data)

$\rightarrow$ social networks are usually degree-assortative

Note that all complex networks are not degree-assortative:

- Internet AS level: $r = -0.215$ (UCLA data)
- human protein network: $r = -0.126$ (Vidal data)
- US power grid network: $r = 0.003$ (Tore Opsahl data)

Structural homophily: other kinds of assortativity

Degree-based assortativity

$$r = \frac{\sum_{j,k} j.k.(e_{jk} - q_j q_k)}{\sigma_q^2} \text{ with } \sigma_q^2 = \frac{1}{n}\sum_k q_k (k - \overline{k})^2$$

r: ratio covariance across ties / variance of remaining degree k

General assortativity

$$\rho = \frac{\sum_{\lambda,\kappa} \lambda.\kappa.(\epsilon_{\lambda\kappa} - \chi_\lambda \chi_\kappa)}{\sigma_\chi^2} \text{ with } \sigma_\chi^2 = \frac{1}{n}\sum_\kappa \chi_\kappa (\kappa - \overline{\kappa})^2$$

ρ: ratio covariance across links / variance of characteristic κ
ϵ: joint probability across link of characteristic κ
χ: probability distribution of characteristic κ

examples for κ: age, salary, number of children. . .

Structural homophily: other kinds of assortativity

Degree-based assortativity

$$r = \frac{\sum_{j,k} j.k.(e_{jk} - q_j q_k)}{\sigma_q^2} \text{ with } \sigma_q^2 = \frac{1}{n}\sum_k q_k(k-\overline{k})^2$$

r: ratio covariance across ties / variance of remaining degree k

General assortativity

$$\rho = \frac{\sum_{\lambda,\kappa} \lambda.\kappa.(\epsilon_{\lambda\kappa} - \chi_\lambda \chi_\kappa)}{\sigma_\chi^2} \text{ with } \sigma_\chi^2 = \frac{1}{n}\sum_\kappa \chi_\kappa(\kappa-\overline{\kappa})^2$$

ρ: ratio covariance across links / variance of characteristic κ
ϵ: joint probability across link of characteristic κ
χ: probability distribution of characteristic κ

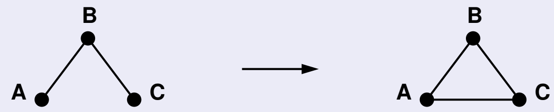
examples for κ: age, salary, number of children...

Transitivity and clustering

Reminder: social networks have a high average clustering

Triadic closure phenomenon

Old concept in sociology (Simmel, 1908).
Hypothesis on the growth dynamics of a network:



Consequences:

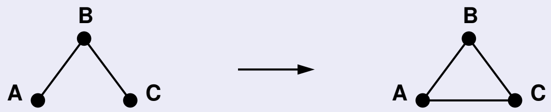
- high clustering
- large number of cliques (complete subgraphs)
- densification over time

Transitivity and clustering

Reminder: social networks have a high average clustering

Triadic closure phenomenon

Old concept in sociology (Simmel, 1908).
Hypothesis on the growth dynamics of a network:

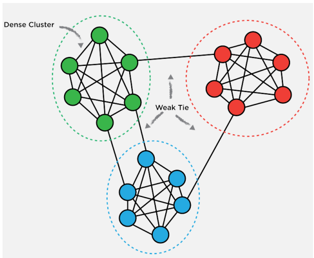


Consequences:

- high clustering
- large number of cliques (complete subgraphs)
- densification over time

Transitivity and clustering

$\Rightarrow$ schematic picture of social networks

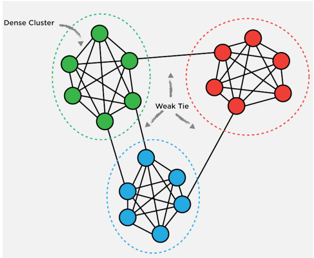


credits image: V.Gauthier

Disclaimer: only a schematic representation
misses overlaps in clusters, groups hierarchy, core/periphery...

Transitivity and clustering

⇒ schematic picture of social networks



credits image: V.Gauthier

Disclaimer: only a schematic representation
misses overlaps in clusters, groups hierarchy, core/periphery...

About weak ties ...

Hypothesis of the strength of weak tie (link) Granovetter - *1973*

A "weak tie" is a link in a social network which represents a relation which is not frequently maintained

It is argued that weak ties play an essential role as they ensure connections between groups

What is a measure of strength of a tie in social sciences?

- frequent contacts
- strong affinity (if measurable)
- structural criterion: many mutual neighbors

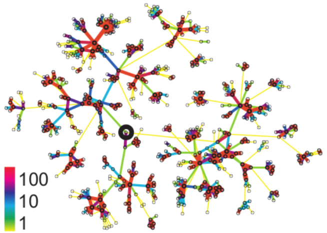
About weak ties ...

Onnela et al. - *2007*

Experimental validation on a phonecall network
→ are weaker links between clusters?

- strength of a relationship = cumulative duration of calls
- link between groups measured with link betweenness

weight (color) = cumulative duration of calls



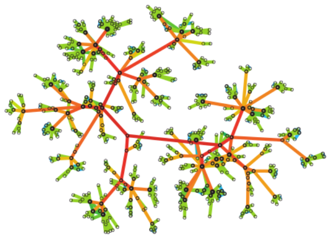
About weak ties ...

Onnela et al. - *2007*

Experimental validation on a phonecall network
→ are weaker links between clusters?

- strength of a relationship = cumulative duration of calls
- link between groups measured with link betweenness

weight (color) = link betweenness



About weak ties . . .

Onnela et al. - *2007*

Experimental validation on a phonecall network

$\rightarrow$ are weaker links between clusters?

Quiz: what could you plot to check the correlation ?

Applying our tools to Social Network Analysis
Epidemic spreading models on graphs
The homophily phenomenon
Local density and community structure
Impact on social contagion

About weak ties . . .

Onnela et al. - *2007*

Experimental validation on a phonecall network

$\rightarrow$ are weaker links between clusters?

Quiz: what could you plot to check the correlation ?

e.g.: link strength vs. link betweenness, correlation coefficient

Applying our tools to Social Network Analysis
Epidemic spreading models on graphs
The homophily phenomenon
Local density and community structure
Impact on social contagion

Effects on spreading in a social network

Examples: innovation spreading, rumor spreading, adversiting. . .

What can we expect from the previous observations?

- fast spreading within a community
- use of weak links to spread from a group to another

In practice hard to measure experimentally:

- "contagion" hard to track and isolate
- spreading rarely reaches a large part of a network

$\rightarrow$ active field of research

Applying our tools to Social Network Analysis
Epidemic spreading models on graphs
The homophily phenomenon
Local density and community structure
Impact on social contagion

Effects on spreading in a social network

Examples: innovation spreading, rumor spreading, adversiting. . .

What can we expect from the previous observations?

- fast spreading within a community
- use of weak links to spread from a group to another

In practice hard to measure experimentally:

- "contagion" hard to track and isolate
- spreading rarely reaches a large part of a network

$\rightarrow$ active field of research

Effects on spreading in a social network

Examples: innovation spreading, rumor spreading, adversiting...

What can we expect from the previous observations?

- fast spreading within a community
- use of weak links to spread from a group to another

In practice hard to measure experimentally:

- "contagion" hard to track and isolate
- spreading rarely reaches a large part of a network

$\rightarrow$ active field of research

Outline

From social networks to epidemic spreading

In SNA, innovation spreading dates back to the 50s

In parallel, epidemic modeling developed

Late 90s, data availability $\Rightarrow$ take into account the social network that supports the spreading

Networks and epidemic models - *Keeling and Eames, 2005*

Traditional (simplified) approach of epidemic model

Compartmental models

Basic assumption:

- random mixing hypothesis
  each individual has an equal chance to come into contact with anyone
- $\Rightarrow$ homogeneous description of individual behaviors

Infection and population complexity $\rightarrow$ compartments

- **S**: susceptible
- **I**: infected
- **R**: recovered
- **E**: exposed
- **M**: maternally-immune ...

Traditional (simplified) approach of epidemic model

Compartmental models

Basic assumption:

- random mixing hypothesis
  each individual has an equal chance to come into contact with anyone
- $\Rightarrow$ homogeneous description of individual behaviors

Infection and population complexity $\rightarrow$ compartments

- **S**: susceptible
- **I**: infected
- **R**: recovered
- **E**: exposed
- **M**: maternally-immune . . .

Traditional (simplified) approach of epidemic model

Equations

Model: set of transition rules from a compartment to another

- when S encounters I, uniform probability of contamination
- after being contaminated, I has uniform probability of recovery
- etc.

Example: SIR model

$$\begin{cases} \frac{dS}{dt} = -\beta \frac{IS}{N} \\ \frac{dI}{dt} = +\beta \frac{IS}{N} - \gamma I \\ \frac{dR}{dt} = +\gamma I \end{cases}$$

Traditional (simplified) approach of epidemic model

Example: SIR model

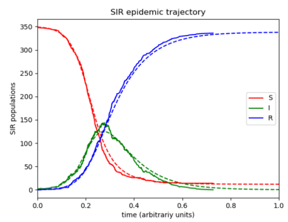
$$\begin{cases} \frac{dS}{dt} = -\beta \frac{IS}{N} \\ \frac{dI}{dt} = +\beta \frac{IS}{N} - \gamma I \\ \frac{dR}{dt} = +\gamma I \end{cases}$$

- $N = S + I + R$: size of the population
- β : infection rate (parameter)
- γ : recovery rate (parameter)

Traditional (simplified) approach of epidemic model

Example: SIR model

$$\begin{cases} \frac{dS}{dt} = -\beta \frac{IS}{N} \\ \frac{dI}{dt} = +\beta \frac{IS}{N} - \gamma I \\ \frac{dR}{dt} = +\gamma I \end{cases}$$



Classic models

- SIR:
  - 3 compartments S, I, R
  - used for disease with lifelong immunity
    ex: measles (rougeole) , whooping cough (coqueluche)
- SIS:
  - 2 compartments S, I
  - used for disease with possible reinfections
    ex: STD such as chlamydia
  - equations

$$\begin{cases} \frac{dS}{dt} = -\beta\frac{IS}{N} + \gamma I \\ \frac{dI}{dt} = +\beta\frac{IS}{N} - \gamma I \end{cases}$$

- SEIS, SEIR, SEIRS, MSEIR, ...

A few useful concepts in epidemiology

- Basic reproductive number R_0: expected number of new infections from a single infection if everyone is susceptible
  ex: for SIR with random mixing, $R_0 = \frac{\beta}{\gamma}$
  some estimated values (without intervention):
  - *measles: 12–18*
  - *seasonal flu: 1–2*
  - *covid-19: 3.3–5.7*
- k value: shape parameter, dispersion parameter, related to the inverse of the dispersion
  random mixing ⇒ homogeneous behavior ⇒ high values of k
  some estimated values (without intervention):
  - *measles: 0.22*
  - *seasonal flu: 2 – 50*
  - *covid-19: 0.16*

A few useful concepts in epidemiology

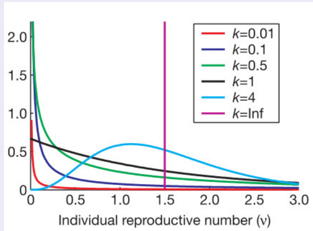
- Basic reproductive number R_0: expected number of new infections from a single infection if everyone is susceptible
  ex: for SIR with random mixing, $R_0 = \frac{\beta}{\gamma}$
  some estimated values (without intervention):
  - *measles: 12–18*
  - *seasonal flu: 1–2*
  - *covid-19: 3.3–5.7*
- k value: shape parameter, dispersion parameter, related to the inverse of the dispersion
  random mixing ⇒ homogeneous behavior ⇒ high values of k
  some estimated values (without intervention):
  - *measles: 0.22*
  - *seasonal flu: 2 – 50*
  - *covid-19: 0.16*

More about the k value

Superspreading and the effect of individual variation on disease emergence -
Lloyd-Smith et al., 2005

Origin

Model for the individual reproductive number distribution: negative binomial distribution

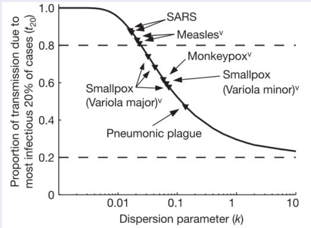


Average $\mu = R_0$, variance $\mu + \frac{\mu^2}{k}$

More about the k value

Superspreading and the effect of individual variation on disease emergence -
Lloyd-Smith et al., 2005

Another interpretation



∗ error of a factor 10 on the X-axis range

More about the importance of k for Covid-19 pandemic:
`https://www.theatlantic.com/health/archive/2020/09/k-overlooked-variable-driving-pandemic/616548/`

Applying our tools to Social Network Analysis
Epidemic spreading models on graphs
Compartmental models in epidemiology
What networks bring to the models

What networks bring

Random mixing assumption implies uniform contact patterns
we know it is not true

Network-based models

- keep the compartments (S, I, R, E, . . .)
- spreading occurs on the contact network
- at each step, nodes may change compartment:
  - I may contaminate S (or E) neighbors
  - I may recover and turns R
  - . . .
- more details: *see practical work*

Applying our tools to Social Network Analysis
Epidemic spreading models on graphs
Compartmental models in epidemiology
What networks bring to the models

What networks bring

Random mixing assumption implies uniform contact patterns
we know it is not true

Network-based models

- keep the compartments (S, I, R, E, . . .)
- spreading occurs on the contact network
- at each step, nodes may change compartment:
  - I may contaminate S (or E) neighbors
  - I may recover and turns R
  - . . .
- more details: *see practical work*

Applying our tools to Social Network Analysis
Epidemic spreading models on graphs
Compartmental models in epidemiology
What networks bring to the models

What networks bring

Random mixing assumption implies uniform contact patterns
we know it is not true

Network-based models

- keep the compartments (S, I, R, E, . . .)
- spreading occurs on the contact network
- at each step, nodes may change compartment:
  - I may contaminate S (or E) neighbors
  - I may recover and turns R
  - . . .
- more details: *see practical work*

Data collection and issues

Network-based epidemiology develops because of data
but how available and reliable are the data?

What is the meaning of an edge?

- *potentially* infectious contact
  disease specific: relatively clear for some diseases (STDs), but airborne diseases?
  ex: TousAntiCovid definition of a contact
- $\Rightarrow$ some degree of arbitrariness

Data collection and issues

Network-based epidemiology develops because of data
but how available and reliable are the data?

What is the meaning of an edge?

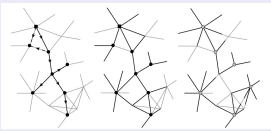
- *potentially* infectious contact
  disease specific: relatively clear for some diseases (STDs), but airborne diseases?
  ex: TousAntiCovid definition of a contact
- $\Rightarrow$ some degree of arbitrariness

Data collection and issues

Network-based epidemiology develops because of data
but how available and reliable are the data?

What are the data collection methods?

- infection tracing: look for infectious individuals that have transmitted the disease
- contact tracing: interview individuals to collect their potentially infectious contacts
- diary-based: based on day-to-day collection by individuals
  ex: cattle breeders in Europe since the 90s



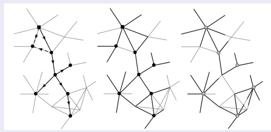
credits image: M.Keeling, K.Eames

Data collection and issues

Network-based epidemiology develops because of data
but how available and reliable are the data?

Issues with data collection methods

- infection tracing: focus on infectious contacts not all contacts, costly
- contact tracing: individual bias, sensitive data, subjective evaluation of danger, heterogeneity of data, costly
- diary-based: individual bias, sensitive data, disconnected network, heterogeneity of data



credits image: M.Keeling, K.Eames

Modeling on artificial networks

Difficult data collection
⇒ large use of network models

A few key-results in this field:

- "shortcuts" (e.g. air connections) break the locality of spreading and geographic wave-like patterns
- hubs (superspreaders) have a dramatic effect on an epidemic, can re-ignite the spreading
- large variability of spreading simulations: heterogeneity of networks ⇒ fluctuations

Applying our tools to Social Network Analysis
Epidemic spreading models on graphs
Compartmental models in epidemiology
What networks bring to the models

Modeling on artificial networks

Difficult data collection
⇒ large use of network models

A few key-results in this field:

- "shortcuts" (e.g. air connections) break the locality of spreading and geographic wave-like patterns
- hubs (superspreaders) have a dramatic effect on an epidemic, can re-ignite the spreading
- large variability of spreading simulations: heterogeneity of networks ⇒ fluctuations

18/18

Applying our tools to Social Network Analysis
Epidemic spreading models on graphs
Compartmental models in epidemiology
What networks bring to the models

Modeling on artificial networks

Difficult data collection
⇒ large use of network models

A few key-results in this field:

- "shortcuts" (e.g. air connections) break the locality of spreading and geographic wave-like patterns
- hubs (superspreaders) have a dramatic effect on an epidemic, can re-ignite the spreading
- large variability of spreading simulations: heterogeneity of networks ⇒ fluctuations

18/18

Applying our tools to Social Network Analysis
Epidemic spreading models on graphs
Compartmental models in epidemiology
What networks bring to the models

Modeling on artificial networks

Difficult data collection
⇒ large use of network models

A few key-results in this field:

- "shortcuts" (e.g. air connections) break the locality of spreading and geographic wave-like patterns
- hubs (superspreaders) have a dramatic effect on an epidemic, can re-ignite the spreading
- large variability of spreading simulations: heterogeneity of networks ⇒ fluctuations

18/18