Network

2017 Big Data Summer Institute

Zhenke Wu[1]

June 22, 2017

[1]Assistant Professor of Biostatistics, U of Michigan, Ann Arbor

Question for Today

- "Game of Thrones: Who is the protagonist?" (Beveridge and Shan 2016, Math Horizons)
- "Why are my friends more popular than me?" (application to early detection of contagious outbreaks: Christakis and Fowler, 2010, PLoS One)

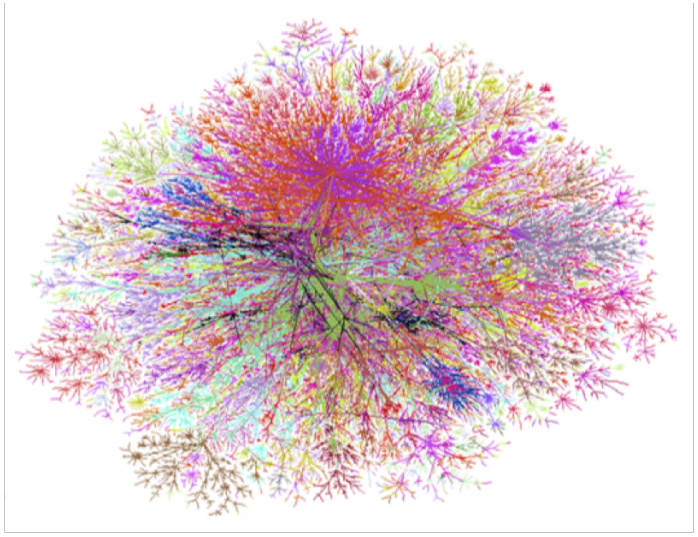
Outline

- Examples and Notations
- Why study networks?
  - Network topology
  - Observations sampled from networks
- What are the common quantitative methods? (not much today)
- References

Examples

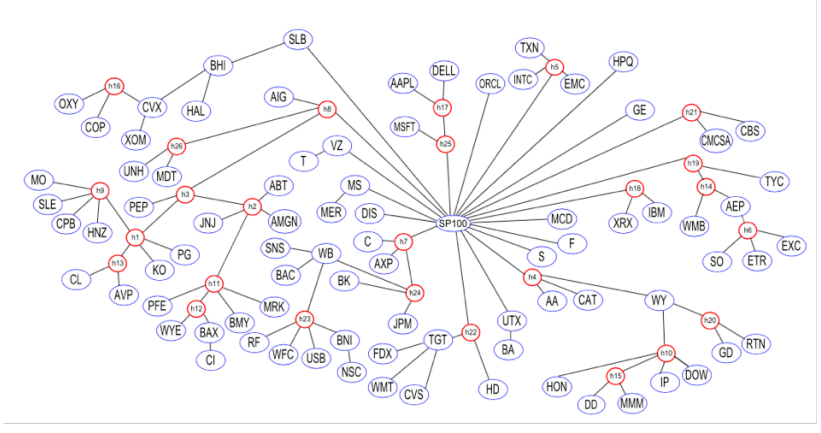
- One of many classifications:
  - Social networks (e.g., Twitter, Facebook, WeChat; Friend formation)
  - Information networks (e.g., World Wide Web)
  - Biological networks (e.g., gene-gene interaction network, human brain functional connnection network, disease transmission in a network)
  - Trade network between companies/countries
  - ...

Examples of Networks



Internet: Bill Cheswick
http://www.cheswick.com/ches/map/gallery/index.html

Airline Network: Northwest Airlines WorldTraveler Magazine



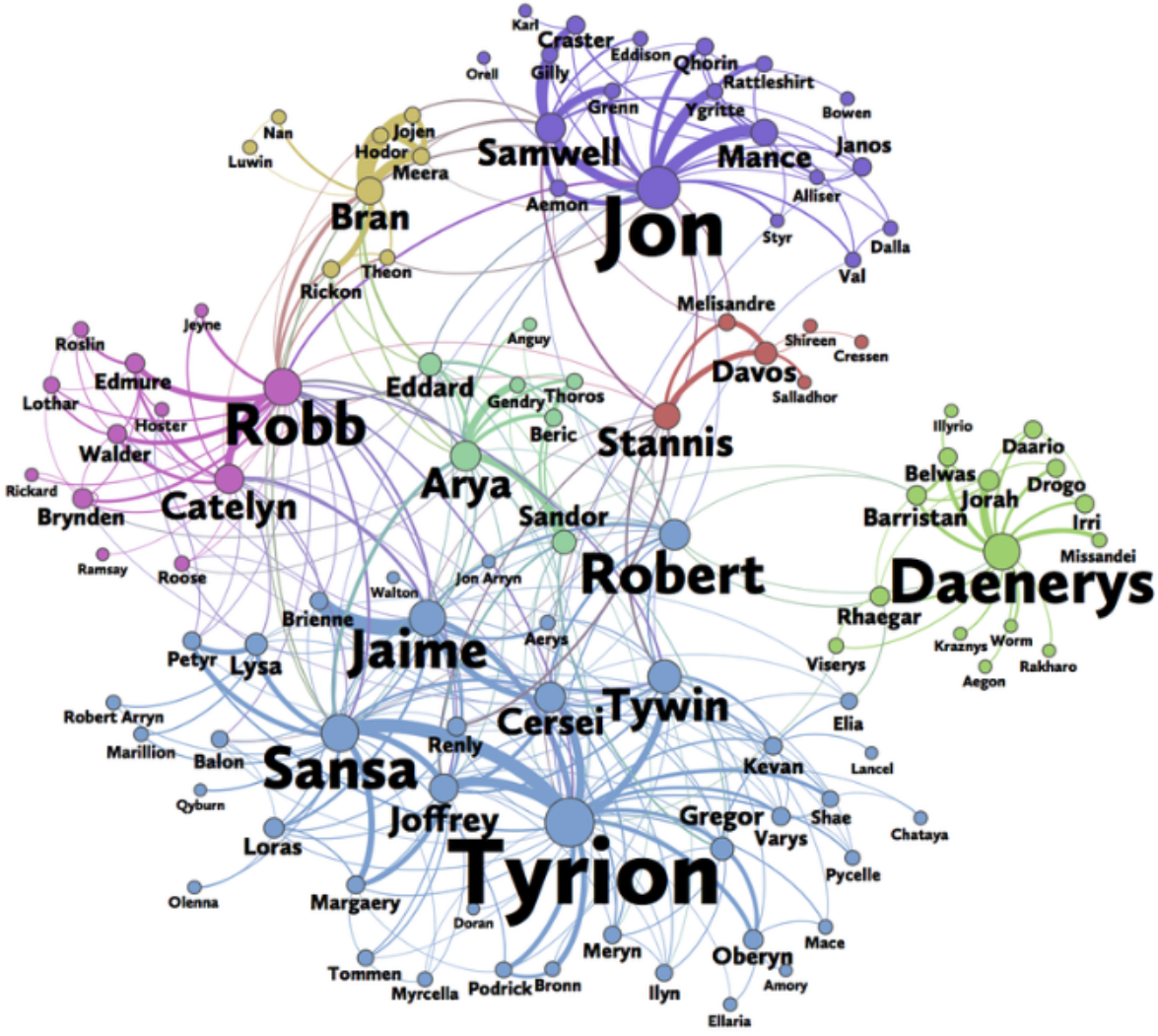
Anandkumar and Valluvan (2013) Annals of Statistics. Figure: Tree graph learned on S&P100 monthly stock return data

New York City Subway. http://web.mta.info/maps/submap.html

Part I: Network of Thrones (Beveridge and Shan, 2016)

Figure 1. The *Game of Thrones* world: Westeros, the Narrow Sea, and Essos (from left to right). Sigils represent the locations of the noble houses at the beginning of the saga.

Part I: Network of Thrones (Beveridge and Shan, 2016)



The color indicates the community, the size of the vertex shows the PageRank, the size of the label shows the betweenness, and a link's thickness shows its weight.

Andrew Beveridge and Jie Shan/Courtesy of Andrew Beveridge/MAA

Game of Thrones Social Network

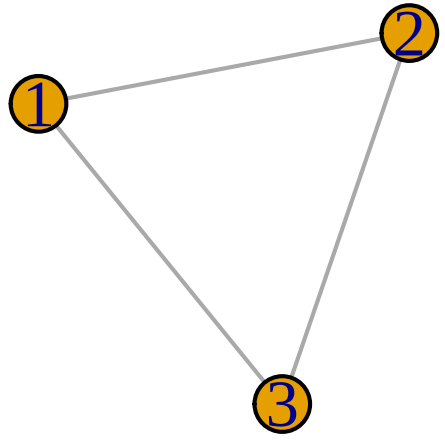
- The third book: **A Storm of Swords**
- 107 characters: ladies, lords, guards, mercenaries, concilmen, consorts, villagers and savages
- Parsed the ebook, assign an edge if two characters appeared within 15 words of one another
- 353 integer-weighted edges: higher weights for stronger relationships (weight $=$ # of co-appearence within 15 words)
- Edge does not necessarily mean friendship; Instead, interaction or were mentioned together.

Questions

- **Community detection**: What are the communities? (Lannisters and King's Landing, Robb's army, Bran and friends, Arya and companions, Jon Snow and the far North, Stannis's forces, and Daerenys and the exotic people of Essos)
- **Protagonist?**

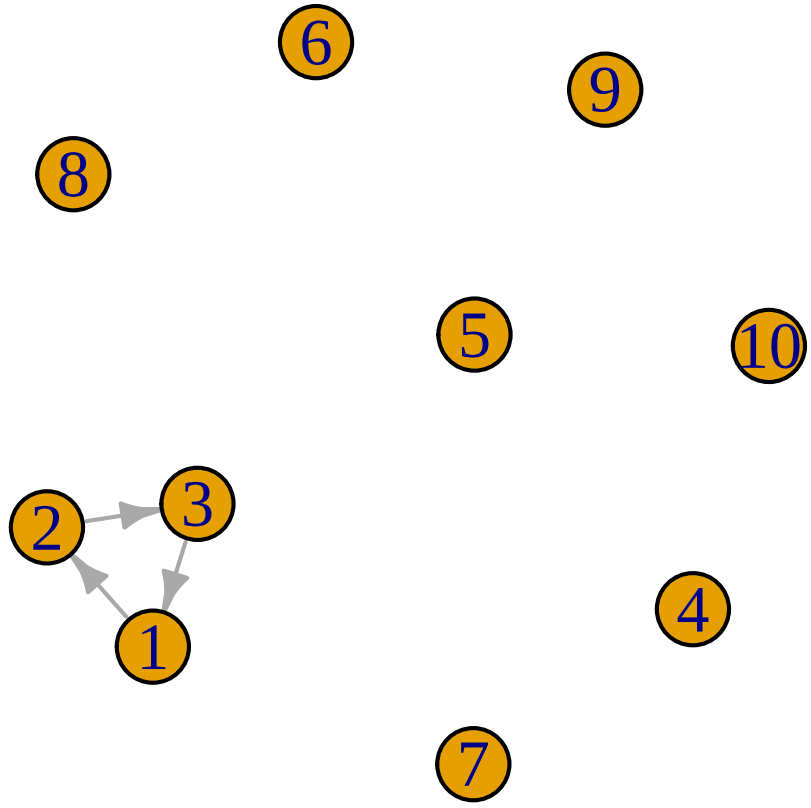
Network Examples

```
# Need library(igraph);
# library(igraphdata) An undirected graph
# with 3 edges:
g1 <- graph(edges = c(1, 2, 2, 3, 3, 1),
    n = 3, directed = F)
plot(g1, vertex.size = 30)
```



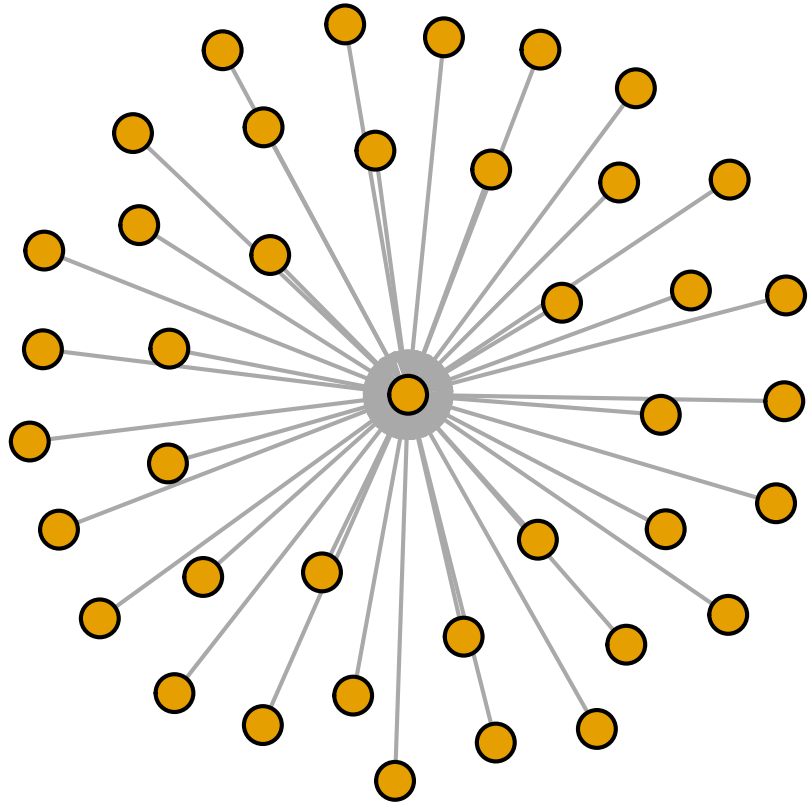
Network Examples

```
# now with 10 vertices, and directed by
# default
g2 <- graph(edges = c(1, 2, 2, 3, 3, 1),
    n = 10)
plot(g2, vertex.size = 20, edge.arrow.size = 0.5)
```



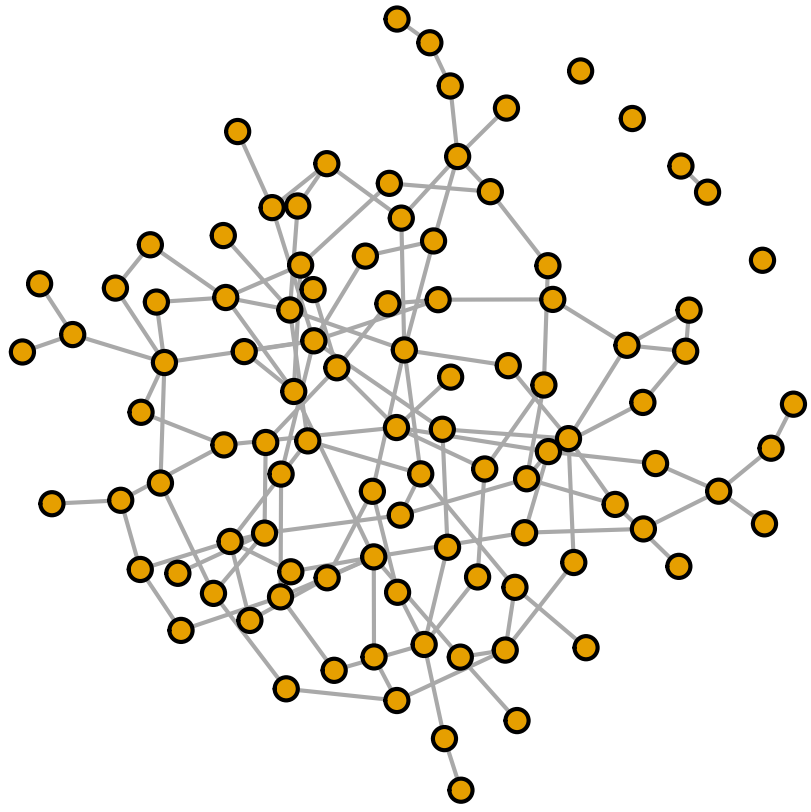
Network Examples (Star Graph)

```
# Star graph
st <- make_star(40)
plot(st, vertex.size = 10, vertex.label = NA,
    edge.arrow.size = 0.3)
```



Network Examples (Erdos-Renyi Model)

```
# Erdos-Renyi Random Graph Model with
# G(n,p) specification
erg <- sample_gnp(n = 100, p = 0.03)
plot(erg, vertex.size = 6, vertex.label = NA)
```



General Themes:

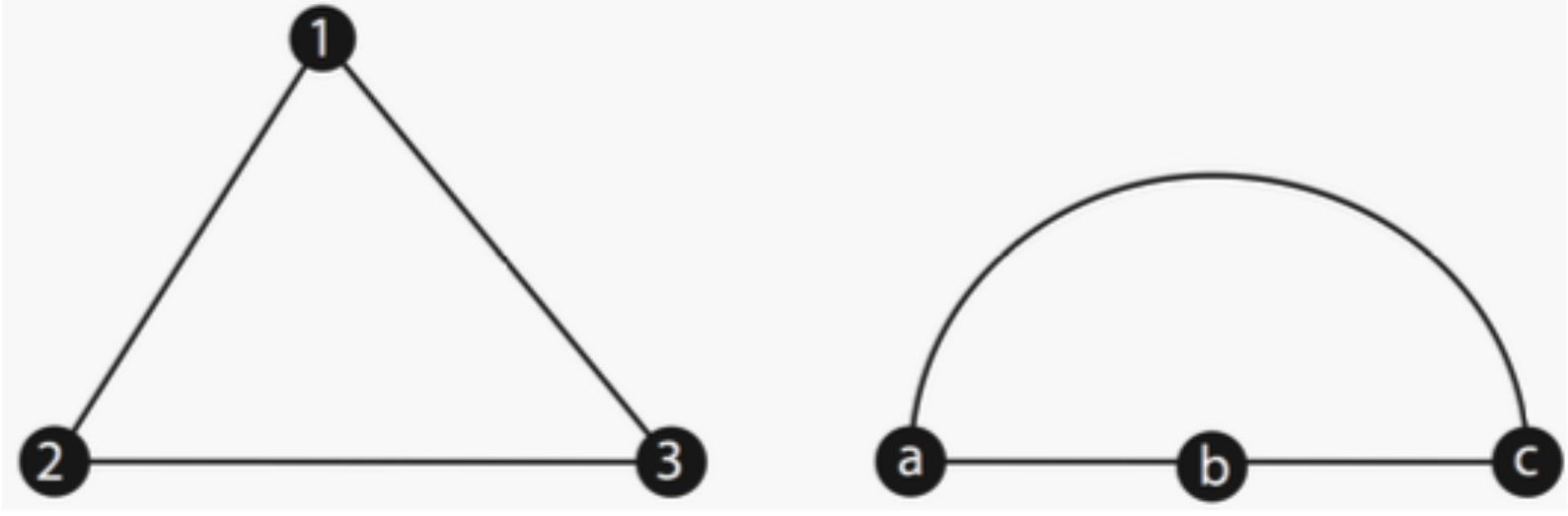
- Formulate mathematical models for observed network patterns and phenomena
- Reason about the model's broader implications about networks, e.g., behavior, population-level dynamics, etc.
- Develop common analytic tools for network data obtained from a variety of settings

Basics

- Network is a graph
- Graphs
  - Mathematical models of network structure
  - Graph: Vertices/Nodes+Edges/Ties/Links
  - A way of specifying relationships among a collection of items

- Graph: Ordered pair $G = (V, E)$
- $V(G)$: vertex set; $E(G)$: edge set
- The vertex pairs may be ordered or unordered, corresponding to directed and undirected graphs
- Some vertex pairs are connected by an edge, some are not
- Two connected vertices are said to be (nearest) neighbors

- Two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are equal if they have equal vertex sets and equal edge sets, i.e., if $V_1 = V_2$ and $E_1 = E_2$ (Note: equality of graph is defined in terms of equality of sets)
- Two graph diagrams (visualizations) are equal if they represent equal vertex sets and equal edge sets



- Edges, depending on context, can signify a variety of things
- Common interpretations
  - Structural connections
  - Interactions
  - Relationships
  - Dependencies
- Often more than one interpretation may be appropriate

- The **degree** of a node in a graph is the number of edges connected to it
- We use d_i to denote the degree of node i
- M edges, then there are $2M$ ends of edges; Also the sum of degrees of all the nodes in the graph: $\sum_i d_i = 2M$
- Nodes in directed graph have **in-degree** and **out-degree**

Link Density

- Consider an undirected network with N nodes
- How many edges can the network have at most?
  - The number of ways of choosing 2 vertices out of N: $N(N-1)/2$
- A graph is fully connected if every possible edge is present

- Let M be the number of edges
- **Link density**: the fraction of edges present, and is denoted by ρ

$$\rho = \frac{2M}{N(N-1)}$$

- Link density lies in $[0, 1]$
- Most real networks have very low ρ
- Dense network: $\rho \to \textit{constant}$ as $N \to \infty$
- Sparse network: $\rho \to 0$ as $N \to \infty$

- An **adjacency matrix** is an $N \times N$ matrix $\mathbf{A}$ where A_{ij} encodes information about the edge between nodes i and j

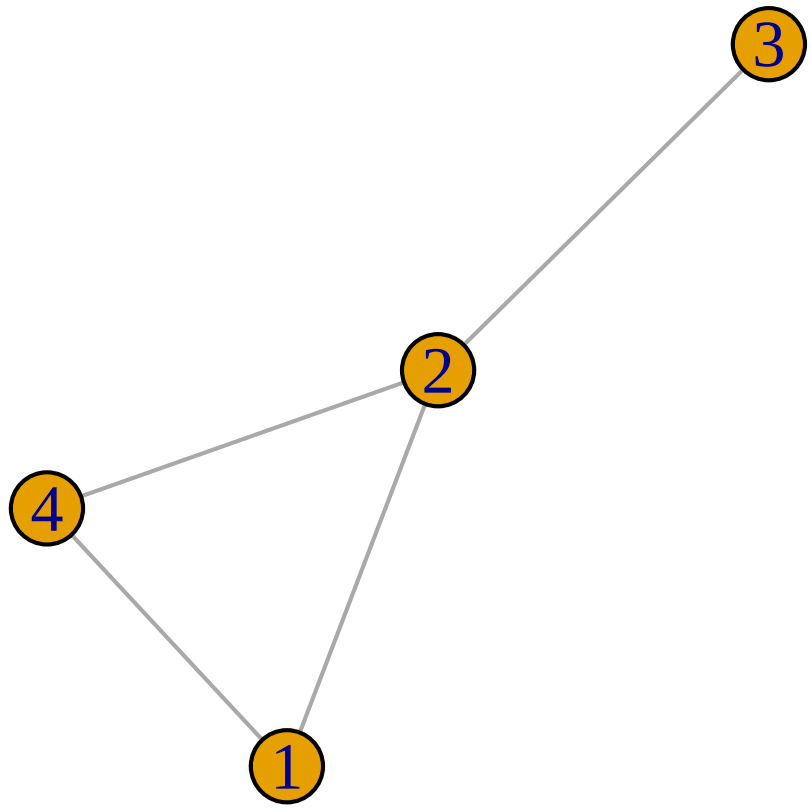
$$\text{e.g. } \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

- **Weighted networks** have weights, covariates, or strength associated with the ties

$$\mathbf{A} = \begin{bmatrix} 0 & .5 & 0 & 2 \\ .5 & 0 & 9 & 3 \\ 0 & 9 & 0 & 0 \\ 2 & 3 & 0 & 0 \end{bmatrix}$$

Network Examples: Adjacency Matrix

```
g_adj <- graph(edges = c(1, 2, 2, 4, 4, 1,
    3, 2), n = 3, directed = FALSE)  # now with 4 vertices
plot(g_adj, vertex.size = 20)
```



Network Examples: Adjacency Matrix

```
A <- get.adjacency(g_adj, sparse = FALSE)
print(A)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    0    1    0    1
## [2,]    1    0    1    1
## [3,]    0    1    0    0
## [4,]    1    1    0    0
```

```
print(A %*% A)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    2    1    1    1
## [2,]    1    3    0    1
## [3,]    1    0    1    1
## [4,]    1    1    1    2
```

The walks of length r are given by $\mathbf{A}^r$;

(Note: walks are different from paths; the former may have multiple identical edges.)

- The shortest between i and j is the *geodesic path*
- How to find its length? (The smallest r such that $[\mathbf{A}^r]_{i,j} > 0$)

Community Detection

- **Community**: roughly speaking, a group of nodes that are more densely connected to each other than to the rest of the network
- One common algorithm: maximizing **modularity**

Community Detection (continued)

- **Modularity**: compare our given network to a network with the same degrees, but in which all edges are rewired at random.
- Global measure
- $d_i = \sum_{j \in V} A_{ij}$
- Suppose i and j belong to community C
- Expected number of randomly rewired edges between i and j: $d_i \frac{d_j}{2M}$, where M is the total $\#$ of edges
- Sum over all vertices in community C: $\sum_{i,j \in C} (A_{ij} - \frac{d_i d_j}{2M})$; Non-negative for a true community

Community Detection (continued) - **Modularity**:

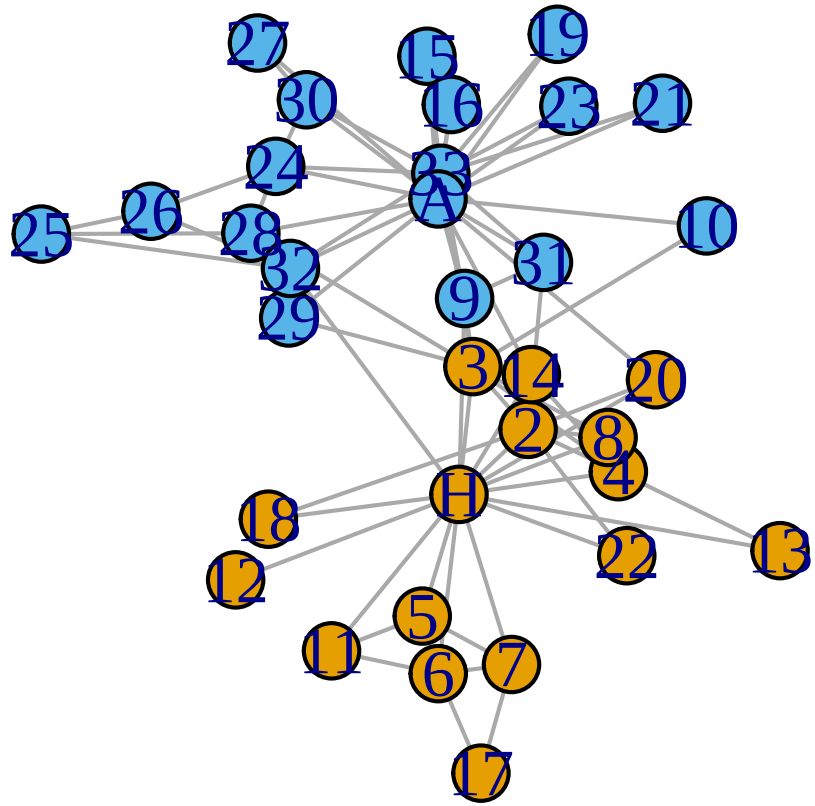
- For a partition $C_1, \ldots, C_L$ of the entire vertex set $V = \cup_\ell C_\ell$:

$$Q = \frac{1}{2M} \sum_{\ell=1}^{L} \sum_{i,j \in C_\ell} \left[A_{ij} - \frac{d_i d_j}{2M} \right]$$

- Maximize Q over all possible partitions $\{C_1, \ldots, C_L\}$ (Louvain method; L need not be prespecified)
- **Result**: The King's Landing community accounts for 37% of the network. [Return to the GoT Network]

Zachary's Karate Club data

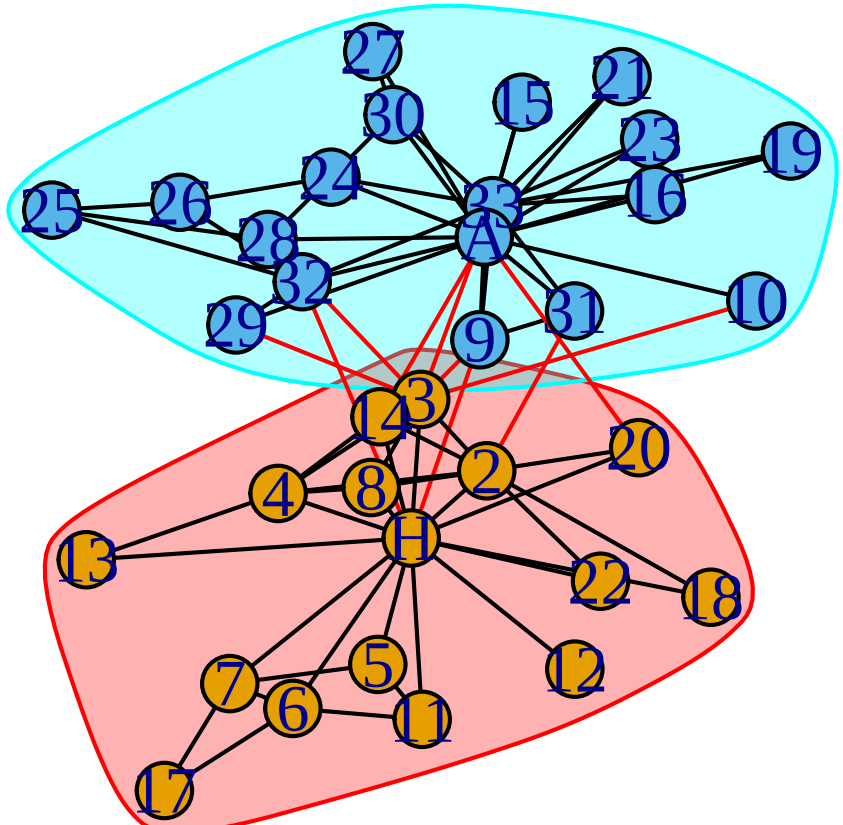
```
data(karate)
# summary(karate)
plot(karate)
```




```
# Actual factions: 1 led by 'Mr Hi', 2
# led by 'John A':
```

Zachary's Karate Club data (continued)

```
# ?communities # check methods.  Fast
# greedy modularity-based clustering
cfg <- cluster_fast_greedy(karate)
# specifying the number of clusters:
plot(structure(list(membership = cutat(cfg,
    2)), class = "communities"), karate)
```



Who's the protagonist?

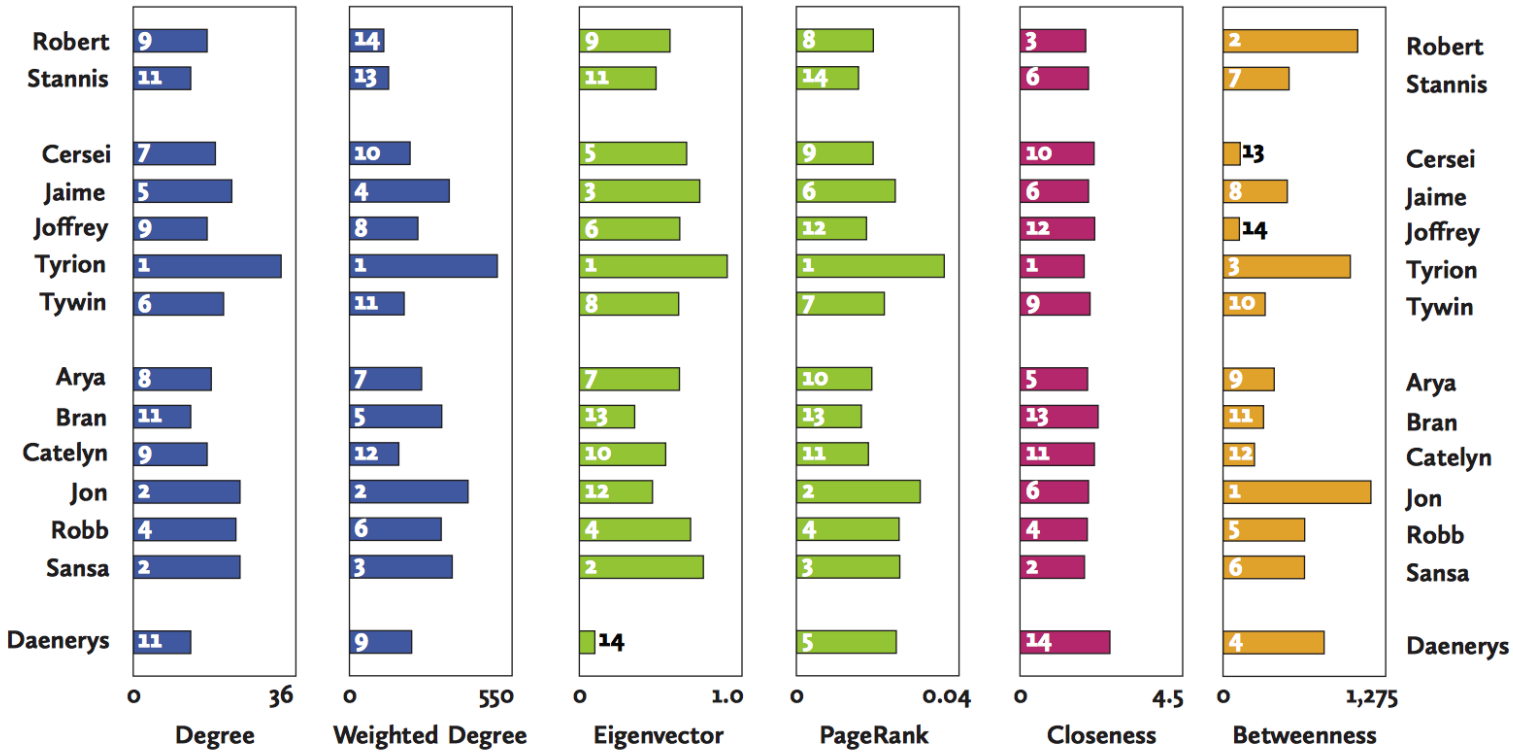


Figure 3. Centrality measures for the network. Larger values correspond to greater importance, except for closeness centrality, where smaller values are better. Numbers in the bars give the rankings of these characters.

Six Concepts of “Centrality"

- **Centrality**: measures how central or important the nodes are in the network
- Proposing new centrality measures and developing algorithms to calculate them is an active field of research

Degree Centrality

- The number of edges incident with the given vertex
- Measures the number of connections to other characters

Weighted Degree Centrality

- The sum of the weight of the incident edges
- Measures the number of interactions

Eigenvector Centrality

- Gives more centrality to nodes whose neighbors are themselves more central
- "It's more important to be connected to influential neighbors than isolated ones"
- Defined as the weighted sum of its neighboring nodes: $c_i = \sum_{j \in V} A_{ij} c_j$
- Equivalent to solving: $\mathbf{Ac} = \kappa \mathbf{c}$

PageRank

- $$y_i = \alpha \sum_{j \in V} \frac{A_{ji}}{d_j} y_j + \beta$$
- β: inherent importance for each vertex
- Importance from neighbors are divided among its neighbors (How is it different from Eigen-Centrality?)
- $\alpha + \beta = 1$, $\alpha, \beta \geq 0$
- Set $\beta = 0.15$; balance the node's inherent importance and influence from its neighbors

Closeness Centrality

- More global
- Average distance from the vertex to all other vertices
- Lower means greater importance
- $l_i = \frac{1}{N} \sum_j d_{ij}$
  - Usually in a small range
  - highly sensitive to small changes in the network
  - Infinite whenever a network has multiple components

Betweenness Centrality

- More global
- How frequently a vertex lies on the geodesic paths between other pairs of vertices
- Let $g_{st}^i = \mathbf{1}\{\textit{vertex } i \textit{ lies on a geodesic path from } s \textit{ to } t\}$
- n_{st} the number of geodesic paths from s to t
- $c_i = \sum_{s,t} \frac{g_{st}^i}{n_{st}}$
  - "Broker of information"
  - Has potential to be highly influential by inserting themselves into the dealings of other parties
  - "Jon Snow is uniquely positioned in the network, with connections to highborn lords, the Night's Watch militia, and the savage wildlings beyond the Wall."

Part II. "Why are my friends more popular than me?"

- "Most people have fewer friends than their friends have, on average."
- Also known as *Friendship Paradox*
- First observed by sociologist Scott Feld in 1991
- Caused by sampling bias: a popular person has an increased likelihood of being your friend.
- Application to Early Detection of Contagious Outbreaks (Christakis and Fowler, 2010)

Why?

- V: the set of vertices - people in the social network
- E: the set of edges - friendship relations among pairs of people
  - Symmetry assumption: if A is a friend of B, then B is a friend of A
- $d(v)$: the number of edges connected to Vertex v, i.e., Person v has $d(v)$ friends
- *The average number of friends of a random person*?
  - $\mu = \frac{\sum_v \#\ of\ friends\ of\ Person\ v}{Total\ \#\ of\ people} = \frac{2|E|}{|V|}$

Why? (continued)

- *The average number of friends of a random person*?
  - $\mu = \frac{\sum_v \#\ of\ friends\ of\ Person\ v}{Total\ \#\ of\ people} = \frac{2|E|}{|V|}$
- *The average number of friends of a random person's friend*?
  - For each *ordered* friendship (u, v)
  - Person u says v is his/her friend
  - Person v has $d(v)$ friends
  - There are $d(v)$ such (u, v) pairs (fix v, vary u)
  - The total number of *ordered* friendships: $2|E|$
  - We get: $\frac{\sum_v d(v)^2}{2|E|} = \mu + \sigma^2/\mu$, where σ^2 is the variance of degrees $\{d(v) : v \in V\}$
- $\mu + \sigma^2/\mu > \mu$, if $\sigma^2 > 0$!

Early Detection of Contagious Outbreaks (C and F 2010)

- Friends of randomly selected individuals are likely to have higher than average centrality
- Individuals near the center of a social network are likely to be infected sooner during the course of an outbreak, on average, than those at the periphery.
- *Challenge*: Unfortunately, mapping a whole network to identify central individuals who might be monitored for infection is typically very difficult.
- *Solution*: simply monitoring the friends of randomly selected individuals.

Early Detection of Contagious Outbreaks (continued)

- Flu outbreak at Harvard College in late 2009 (Sep 1 - Dec 31)
- 744 students: either members of a group of randomly chosen individuals or a group of their friends
  - 319 random individuals from 6,650 Harvard undergrads
  - 425 "friends": named as a friend at least once by a member of the above random sample
- University Health Services Records (Diagnosis by medical staff): Sep 1 to Dec 31

Results of Early Detection

- Progression of epidemic in the friend group occurred 13.9 days (95% CI 9.9,16.6) in advance of the randomly chosen group
- The friend group showed a significant lead time ($p<0.05$) on day 16 of the epidemic, a full 46 days before the peak in daily incidence in the population
- Implication: could be used to provide additional time to react to epidemics under surveillance

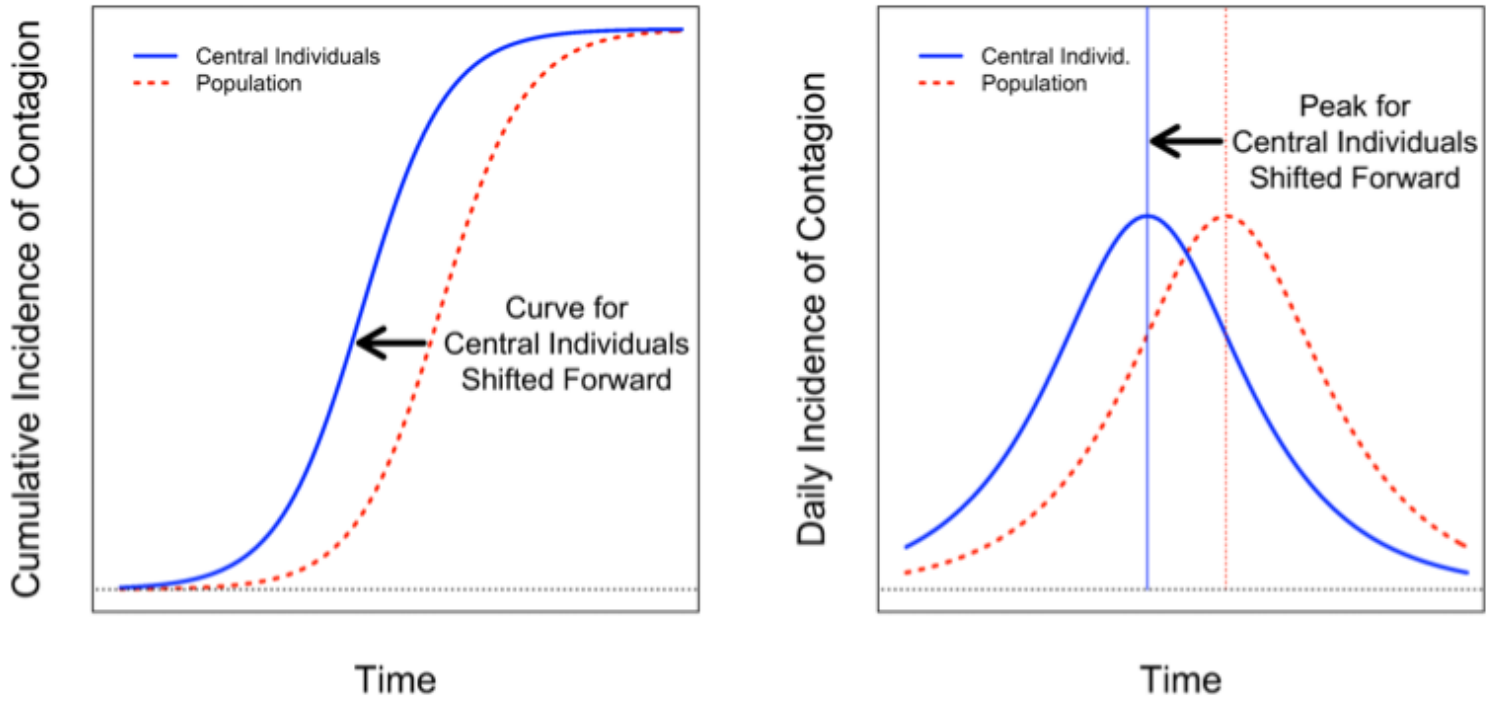


Figure 2. Theoretical expectations of differences in contagion between central individuals and the population as a whole. A contagious process passes through two phases, one in which the number of infected individuals exponentially increases as the contagion spreads, and one in which incidence exponentially decreases as susceptible individuals become increasingly scarce. These dynamics can be modeled by a logistic function. Central individuals lie on more paths in a network compared to the average person in a population and are therefore more likely to be infected early by a contagion that randomly infects some individuals and then spreads from person to person within the network. This shifts the S-shaped logistic cumulative incidence function forward in time for central individuals compared to the population as a whole (left panel). It also shifts the peak infection rate forward (right panel).
doi:10.1371/journal.pone.0012948.g002

Main Points Once Again

- "Game of Thrones: Who is the protagonist?" (centrality)
- "Why are my friends more popular than me?" (network driven sampling bias; could be beneficial)

Did not discuss today

- Generate a random network:
  1. Erdos-Renyi (E-R) model, or E-R random graph named after Hungarian mathematicians; Also known as Poisson random graph (degree distribution of the model follows a Poisson)
  2. Barabasi-Albert model (preferential attachment)
  3. Small-world model/Watts-Strogatz model (high transitiity; small-world property)
  4. Exponential Random Graph Models (ERGM)
  5. Stochastic block models (community structure)
  6. Latent space models

- Network Fundamentals
  1. Basics: Chapter 6; Descriptors: Chapter 7-8; Models: Chapter 12-15, Newman (2010). [Networks: An Introduction. Oxford University Press.]
- Social Networks:
  1. Chapter 3, Newman book.
  2. Hoff, Raftery and Handcock (2002). Latent Space Approaches to Social Network Analysis. *JASA*.
- Social Influence (Peer-Effects; Contagion):
  1. Christakis and Fowler (2007). The Spread of Obesity in a Large Social Network over 32 Years. NEJM.
  2. Responses to CF2007: Cohen-Cole and Fletcher (2008); Lyons (2011); Shalizi and Thomas (2011); and More
  3. O'Malley et al. (2014). Estimating Peer Effects in Longitudinal Dyadic Data Using Instrumental Variables. *Biometrics*.
- Infectious Disease Dynamics
  1. Chapter 21, Easley and Kleinberg (2010). [Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press.]

- Notes partially sourced from Betsy Ogburn and JP Onella