Social Network Analysis: Lecture 3-Network Characteristics

Donglei Du
(ddu@unb.ca)

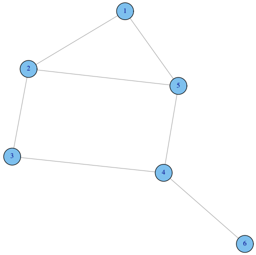
Faculty of Business Administration, University of New Brunswick, NB Canada Fredericton
E3B 9Y2

Table of contents

Network characteristics

- Degree distribution
- Path distribution
- Clustering coefficient distribution
- Size of the giant component
- Community structure
- Assortative mixing (a.k.a., homophily or Heterophily in social network)

Degree distribution for undirected graph



- Degree distribution: A frequency count of the occurrence of each degree.
- First the degrees are listed below:

node	degree
1	2
2	3
3	2
4	3
5	3
6	1

Degree distribution for undirected graph

- The degree distribution therefore is:

degree	frequency
1	1/6
2	2/6
3	3/6

- Average degree: let $N = |V|$ be the number of nodes, and $L = |E|$ be the number of edges:

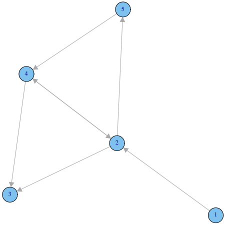
$$\langle K \rangle = \frac{\sum_{i=1}^{n} deg(i)}{N} = \frac{2L}{N}$$

- $\langle k \rangle = 2(7)/6 = 7/3$ for the above graph.

R code based on package igraph: Degree distribution

```
rm(list=ls())# clear memory
library(igraph) # load package igraph
#####################################################################################
#Generate undirected graph object from adjacency matrix
#####################################################################################
adjm_u<-matrix(
   c(0, 1, 0, 0, 1, 0,
1, 0, 1, 0, 1, 0,
0, 1, 0, 1, 0, 0,
0, 0, 1, 0, 1, 1,
1, 1, 0, 1, 0, 0,
0, 0, 0, 1, 0, 0), # the data elements
   nrow=6,              # number of rows
   ncol=6,              # number of columns
   byrow = TRUE)        # fill matrix by rows
g_adj_u <- graph.adjacency(adjm_u, mode="undirected")
# calculate the degree and degree distribution
degree.distribution(g_adj_u)
degree(g_adj_u,loops = FALSE)
```

Degree and degree distribution for directed graph



- Indegree of any node i: the number of nodes destined to i.
- Outdegree of any node i: the number of nodes originated at i.
  - Every loop adds one degree to each of the indegree and outdegree of a node.

node	indegree	outdegree
1	0	1
2	2	3
3	2	0
4	2	2
5	1	1

Degree and degree distribution for directed graph

- Degree distribution: A frequency count of the occurrence of each degree

indegree	frequency	outdegree	frequency
0	1/5	0	1/5
1	1/5	1	2/5
2	3/5	2	1/5
		3	1/5

- Average degree: let $N = |V|$ be the number of nodes, and $L = |E|$ be the number of arcs:

$$\langle K^{in} \rangle = \frac{\sum_{i=1}^{n} deg_{in}(i)}{N} = \frac{\sum_{i=1}^{n} deg_{out}(i)}{N} = \frac{L}{N}$$

- $\langle K^{in} \rangle = \langle K^{out} \rangle = 7/5$ for the above graph.

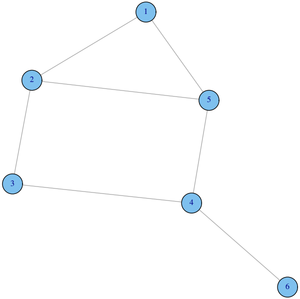
R code based on package igraph: degree

```
rm(list=ls())# clear memory
library(igraph)# load package igraph
##########################################################################################
#Generate directed graph object from adjacency matrix
##########################################################################################
adjm_d<-matrix(
   c(0, 1, 0, 0, 0,
0, 0, 1, 1, 1,
0, 0, 0, 0, 0,
0, 1, 1, 0, 0,
0, 0, 0, 1, 0), # the data elements
   nrow=5,              # number of rows
   ncol=5,              # number of columns
   byrow = TRUE)        # fill matrix by rows
g_adj_d <- graph.adjacency(adjm_d, mode="directed")
# calculate the indegree and outdegree distribution
degree.distribution(g_adj_d, mode="in")
degree.distribution(g_adj_d, mode="out")
degree(g_adj_d,mode="in",loops = FALSE)
degree(g_adj_d,mode="out",loops = FALSE)
```

Why do we care about degree?

- Degree is interesting for several reasons.
  - the simplest, yet very illuminating centrality measure in a network:
    - In a social network, the ones who have connections to many others might have more influence, more access to information, or more prestige than those who have fewer connections.
  - The degree is the immediate risk of a node for catching whatever is flowing through the network (such as a virus, or some information)

Path distance distribution for undirected graph



- Path distribution: A frequency count of the occurrence of each path distance.
- First the path distances are listed below:

	1	2	3	4	5	6
1	0	1	2	2	1	3
2	1	0	1	2	1	3
3	2	1	0	1	2	2
4	2	2	1	0	1	1
5	1	1	2	1	0	2
6	3	3	2	1	2	0

Path distance distribution for undirected graph

- The path distance distribution D therefore is:

distance	frequency
1	7/15
2	6/15
3	2/15

- Average path distance: let $N = |V|$ be the number of nodes:

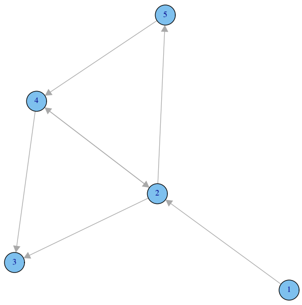
$$\langle D \rangle = \frac{\sum_{i=1}^{n} dist(i,j)}{\binom{N}{2}}$$

- $\langle D \rangle = \mathbb{E}[D] = 5/3$ for the above graph.

R code based on package igraph: Path distribution

```
rm(list=ls())# clear memory
library(igraph) # load package igraph
##########################################################################################
#Generate undirected graph object from adjacency matrix
##########################################################################################
adjm_u<-matrix(
   c(0, 1, 0, 0, 1, 0,
1, 0, 1, 0, 1, 0,
0, 1, 0, 1, 0, 0,
0, 0, 1, 0, 1, 1,
1, 1, 0, 1, 0, 0,
0, 0, 0, 1, 0, 0), # the data elements
   nrow=6,              # number of rows
   ncol=6,              # number of columns
   byrow = TRUE)        # fill matrix by rows
g_adj_u <- graph.adjacency(adjm_u, mode="undirected")
# calculate the path distribution
shortest.paths(g_adj_u)
average.path.length(g_adj_u)
path.length.hist(g_adj_u) # $res is the histogram of distances,
  # $unconnected is the number of pairs for which the first vertex is not
  # reachable from the second.
```

Path distance distribution for directed graph



- Path distribution: A frequency count of the occurrence of each path distance.
- First the path distances are listed below:

	1	2	3	4	5
1	0	1	2	2	2
2	Inf	0	1	1	1
3	Inf	Inf	0	Inf	Inf
4	Inf	1	1	0	2
5	Inf	2	2	1	0

Path distance distribution for directed graph

- The path distance distribution D therefore is:

Distance	Frequency
1	7/13
2	6/13

- Average path distance: let $N = |V|$ be the number of nodes:

$$\langle D \rangle = \frac{\sum_{i<1} dist(i,j)}{\binom{N}{2}}$$

- $\langle D \rangle = \mathbb{E}[D] = 19/13$ for the above graph.

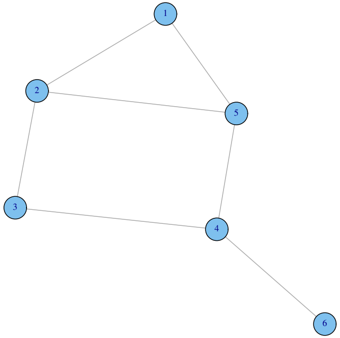
R code based on package igraph: degree

```
rm(list=ls())# clear memory
library(igraph)# load package igraph
###################################################################################
#Generate directed graph object from adjacency matrix
###################################################################################
adjm_d<-matrix(
   c(0, 1, 0, 0, 0,
0, 0, 1, 1, 1,
0, 0, 0, 0, 0,
0, 1, 1, 0, 0,
0, 0, 0, 1, 0), # the data elements
   nrow=5,              # number of rows
   ncol=5,              # number of columns
   byrow = TRUE)        # fill matrix by rows
g_adj_d <- graph.adjacency(adjm_d, mode="directed")
shortest.paths(g_adj_d, mode="out")
shortest.paths(g_adj_d, mode="in")
average.path.length(g_adj_d)
path.length.hist (g_adj_d) # $res is the histogram of distances,
  # $unconnected is the number of pairs for which the first vertex is not
  # reachable from the second.
```

Why do we care about path?

- Path is interesting for several reasons.
  - Path mean connectivity.
  - Path captures the indirect interactions in a network, and individual nodes benefit (or suffer) from indirect relationships because friends might provide access to favors from their friends and information might spread through the links of a network.
  - Path is closely related to small-world phenomenon.
  - Path is related to many centrality measures.
  - ...

Clustering coefficient Distribution for undirected graph



- Recall the definition of local clustering coefficient:

$$\begin{aligned} CC(A) &= \mathbb{P}(B \in N(C) | B, C \in N(A)) \\ &= \mathbb{P}(\text{two randomly selected friends of } A \text{ are friends}) \\ &= \mathbb{P}(\text{fraction of pairs of } A\text{'s friends that are linked to each other}) \\ &= \mathbb{P}(\text{density of the neighboring subgraph}). \end{aligned}$$

- We can also define the global clustering coefficient based on the concept of triplets of nodes.
- A triplet consists of three nodes that are connected by either two (open triplet) or three (closed triplet) undirected ties.
  - A triangle consists of three closed triplets, one centered on each of the nodes.
- The global clustering coefficient is the number of closed triplets (or 3 x triangles) over the total number of triplets (both open and closed):

$$CC = \frac{3 \times \text{number of triangles}}{\text{number of triplets}} = \frac{\text{number of closed triplets}}{\text{number of triplets}}.$$

- Clustering coefficient distribution: A frequency count of the occurrence of each clustering coefficient.
- First the clustering coefficient are listed below:

node	clustering coefficient
1	1
2	1/3
3	0
4	0
5	1/3
6	NaN

Clustering coefficient Distribution for undirected graph

- The Clustering coefficient Distribution therefore is:

Clustering coefficient C	Frequency
0	2/5
1/3	2/5
1	1/5

- Average Clustering coefficient: let $N = |V|$ be the number of nodes:

$$\langle C \rangle = \frac{\sum_{i=1}^{n} CC(I)}{N}$$

- $\langle C \rangle = \mathbb{E}[C] = 1/3$ for the above graph.
- The global clustering coefficient is $3/11 = 0.272727\ldots$
  - First count how many configurations of the form ij, jk there are in the network: 1:1; 2:3; 3:1; 4:3;5:3;6:0. So there are 1+3+1+3+3=11 such congurations in the network.
  - Second count how many triangles there are in the network: there is only one triangle, resulting three closed triplets..

Differences in Clustering Measures

- For the previous example, the average clustering is $1/3$ while the global clustering is $3/11$.
- These two common measures of clustering can differ. Here the average clustering is higher than the overall clustering, it can also go the other way.
- Moreover, it is not hard to generate networks where the two measures can produce very different numbers for the same network.

R code based on package igraph: Clustering coefficient distribution

```
rm(list=ls())# clear memory
library(igraph) # load package igraph
#######################################################################################
#Generate undirected graph object from adjacency matrix
#######################################################################################
adjm_u<-matrix(
   c(0, 1, 0, 0, 1, 0,
1, 0, 1, 0, 1, 0,
0, 1, 0, 1, 0, 0,
0, 0, 1, 0, 1, 1,
1, 1, 0, 1, 0, 0,
0, 0, 0, 1, 0, 0), # the data elements
   nrow=6,              # number of rows
   ncol=6,              # number of columns
   byrow = TRUE)        # fill matrix by rows
g_adj_u <- graph.adjacency(adjm_u, mode="undirected")
# Calculate the clustering coefficient
transitivity(g_adj_u, type="local")# local clustering
transitivity(g_adj_u, type="average") #average clustering
transitivity(g_adj_u)# global clustering: the ratio of the triangles
               # and the connected triples in the graph.
```

Why do we care about clustering coefficient? I

- Clustering is interesting for several reasons.
  - A clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. Evidence suggests that in most real-world networks, and in particular social networks, nodes tend to create tightly knit groups characterized by a relatively high density of ties; this likelihood tends to be greater than the average probability of a tie randomly established between two nodes.
  - Empirically vertices with higher degree having a lower local clustering coefficient on average.
  - Local clustering can be used as a probe for the existence of so-called structural holes in a network, which are missing links between neighbors of a person.

Why do we care about clustering coefficient? II

- Structural holes can be bad when are interested in efficient spread of information or other traffic around a network because they reduce the number of alternative routes information can take through the network.
- Structural holes can be good thing for the central vertex whose friends lack connections because they give i power over information flow between those friends.
- The local clustering coefficient measures how influential i is in this sense, taking lower values the more structural holes there are in the network around i.

- Local clustering can be regarded as a type of centrality measure, albeit one that takes small values for powerful individuals rather than large ones.

The sizes of giant components

- A *giant component* is a connected component (strongly connected component for directed network) in a large network, when its size is a constant fraction of the entire graph.
- Formally, let N_1 be the size of a connected component C in a network of size N, then C is a giant component if

$$\lim_{N\to\infty} \frac{N_1}{N} = c > 0.$$

Community structure

- Network nodes are joined together in tightly knit groups, between which there are only looser connections.
- Refs: (Girvan and Newman, 2002)

Assortative mixing

- Assortative mixing (a.k.a., homophily or Heterophily in social network): the tendency of vertices to connect to others that are alike.

Erdös-Rényi Random Network

- The Erdös-Rényi network (a.k.a. Poisson Metwork) is a random graph $G(N, p)$ with N labeled nodes where each pair of nodes is connected by a preset probability p:
  - Fix node number N.
  - Among all possible edges $\binom{N}{2}$, include each edge with probability p independently.
- N and p do not uniquely define the network: there are $2^{\binom{N}{2}}$ different realizations of it.
- Although the random graph is certainly not a realistic model of most networks, but simple models of networks like this can give us a feel for how more complicated real-world systems should behave in general.
- Let us see some simulation through NetLogo:
  - `http://ccl.northwestern.edu/netlogo/`
  - Go to File/Model Library/Networks: Erdös-Réni Random Model (choose Giant Component)

R code base don package igraph: generating the Erdös-Rényi Random Network

```
>library(igraph)
> g <- erdos.renyi.game(100, 1/100)
> tkplot(g) # interactive plot
```

Simulation of the Erdös-Rényi Random Network through NetLogo

- http://ccl.northwestern.edu/netlogo/
- Go to File/Model Library/Networks/Giant Component

Number of edges distribution for the Erdös-Rényi Random Network I

- If we randomly selected one random graph among all the possible networks: then the probability to have exactly ℓ links in a network of N nodes and probability p:

$$P(L = \ell) = \binom{\binom{N}{2}}{\ell} p^{\ell} (1-p)^{\binom{N}{2}-\ell}.$$

- So the average density is

$$\frac{p\binom{N}{2}}{\binom{N}{2}} = p$$

Number of edges distribution for the Erdös-Rényi Random Network II

- The parameter p in this model can be thought of as a weighting function.
- As p increases from 0 to 1, the model becomes more and more likely to include graphs with more edges and less and less likely to include graphs with fewer edges.
- In particular, the case $p = 0.5$ corresponds to the case where all $2^{\binom{N}{2}}$ graphs on N vertices are chosen with equal probability.

Degree Distribution for the Erdös-Rényi Random Network

Binomial

$\Downarrow$

Approximately **Poisson**

$\Downarrow$

Approximately **Normal**

Degree Distribution for the Erdös-Rényi Random Network is Binomial

- **Binomial**: let K be the degree of a random chosen node, then it can be connected to any of the remaining node independently with probability p, and hence $K \sim B(N-1, p)$:

$$P(K = k) \;=\; C_{N-1}^{k} p^k (1-p)^{N-1-k}.$$

with mean and variance

$$\begin{aligned} \langle K \rangle &= \mathbb{E}[K] = (N-1)p; \\ \sigma^2 &= (N-1)p(1-p). \end{aligned}$$

Degree Distribution for the Erdös-Rényi Random Network is approximately Poisson

- Approximately **Poisson**: $B(N-1, p) \approx P(\lambda)$ with $\lambda = p(N-1) = \langle K \rangle$, for large N and small p (say $N \geq 100$ and $Np \leq 10$)

$$P(K = k) \approx e^{-\langle K \rangle} \frac{\langle K \rangle^k}{k!}, \text{ for large } N \text{ and small } p.$$

with mean and variance all equal to λ.

Degree Distribution for the Erdös-Rényi Random Network is approximately Normal

- Approximately **Normal**: $= N(\lambda, \lambda) \approx P(\lambda)$, for sufficiently large values of λ, (say $\lambda > 1000$; for smaller λ, the continuity correction should be performed):

$$P(K = k) \approx N(\langle K \rangle, \langle K \rangle) \text{ for large } \langle K \rangle.$$

Path distance distribution for the Erdös-Rényi Random Network

- Path distance distribution is hard to find. So we focus on the expectation.
- The average path distance in the random network is approximately

$$\langle L\rangle \approx \frac{\log n}{\log\langle K\rangle}$$

- Idea: Average number of friends at distance d:

$$N_d = \langle K\rangle^d$$

implying that

$$n = \langle K\rangle + \langle K\rangle^1 + \ldots + \langle K\rangle^d \approx \langle K\rangle^d$$

Clustering coefficient distribution for the Erdös-Rényi Random Network

- Clustering coefficient distribution is hard to find. So we focus on the expectation.
- The average Clustering coefficient in the random network is approximately

$$\langle C \rangle \approx \frac{\langle K \rangle}{n}$$

- Randomly select a node i, there are k_i friends, leading to $k_i(k_i - 1)/2$ maximum possible edges, and each will appear with probability p. So the average

$$\langle C \rangle = p \approx \frac{\langle K \rangle}{n}$$

Phase transition of the size of the giant component in the Erdös-Rényi Random Network

- The largest component in the ER random graph has constant size 1 when $p = 0$ and extensive size n when $p = 1$.
- An interesting question to ask is how the transition between these two extremes occurs if we construct random graphs with gradually increasing values of p, starting at 0 and ending up at 1—this is bond percolation!
- It turns out that the size of the largest component undergoes a sudden change, or phase transition, from constant size to extensive size at one particular special value of $p = 1/n$.

The size of the giant component in the Erdös-Rényi Random Network (Bollobás et al., 2001)

- If $p < \frac{1}{n}$
  - with high probability, there is no giant component, with all connected components of the graph having size $O(\log n)$.
- If $p > \frac{1}{n}$
  - with high probability, there is a single giant component, with all other components having size $O(\log n)$.
- If $p = \frac{1}{n}$
  - with high probability, the number of vertices in the largest component of the graph is proportional to $n^{2/3}$.
- See Appendix for an asymptotic analysis Go

Community structure in the Erdös-Rényi Random Network

- Nope!

Assortative mixing in the Erdös-Rényi Random Network

- Nope!

Characteristics of the random network: summary and illustration in Netlogo

- Sparsity: Average density $= p$.
- Degree distribution: Poisson distribution

$$P(K = k) = \binom{n}{k-1} p^k (1-p)^{n-k} \approx e^{-\langle K \rangle} \frac{\langle K \rangle^k}{k!}.$$

- Average path: small world

$$\langle D \rangle \approx \frac{\log n}{\log \langle K \rangle}$$

- Average clustering coefficient: low for large network

$$\langle C \rangle = p \approx \frac{\langle K \rangle}{n}$$

- The threshold for the emergence of the giant component is

$$p = \frac{1}{n} \text{ or } \langle K \rangle \approx 1$$

- No community structure
- No assortative mixing

Network characteristics for real network

- Sparsity: $|E| = O(n)$ edges.
- Degree distribution: Power distribution (scale-free)
- Average path: $O(\log n)$, small world
- Average clustering coefficient: high for large network (compared to random network)
- Giant component: common
- Community structures: common
- Assortative mixing: common

Network characteristics for real networks

	Network	Type	n	m	c	S	ℓ	α	C	C_{WS}	r
Social	Film actors	Undirected	449 913	25 516 482	113.43	0.980	3.48	2.3	0.20	0.78	0.208
	Company directors	Undirected	7 673	55 392	14.44	0.876	4.60	–	0.59	0.88	0.276
	Math coauthorship	Undirected	253 339	496 489	3.92	0.822	7.57	–	0.15	0.34	0.120
	Physics coauthorship	Undirected	52 909	245 300	9.27	0.838	6.19	–	0.45	0.56	0.363
	Biology coauthorship	Undirected	1 520 251	11 803 064	15.53	0.918	4.92	–	0.088	0.60	0.127
	Telephone call graph	Undirected	47 000 000	80 000 000	3.16			2.1			
	Email messages	Directed	59 812	86 300	1.44	0.952	4.95	1.5/2.0		0.16	
	Email address books	Directed	16 881	57 029	3.38	0.590	5.22	–	0.17	0.13	0.092
	Student dating	Undirected	573	477	1.66	0.503	16.01	–	0.005	0.001	−0.029
	Sexual contacts	Undirected	2 810					3.2			
Information	WWW nd.edu	Directed	269 504	1 497 135	5.55	1.000	11.27	2.1/2.4	0.11	0.29	−0.067
	WWW AltaVista	Directed	203 549 046	1 466 000 000	7.20	0.914	16.18	2.1/2.7			
	Citation network	Directed	783 339	6 716 198	8.57			3.0/–			
	Roget's Thesaurus	Directed	1 022	5 103	4.99	0.977	4.87	–	0.13	0.15	0.157
	Word co-occurrence	Undirected	460 902	16 100 000	66.96	1.000		2.7		0.44	
Technological	Internet	Undirected	10 697	31 992	5.98	1.000	3.31	2.5	0.035	0.39	−0.189
	Power grid	Undirected	4 941	6 594	2.67	1.000	18.99	–	0.10	0.080	−0.003
	Train routes	Undirected	587	19 603	66.79	1.000	2.16	–		0.69	−0.033
	Software packages	Directed	1 439	1 723	1.20	0.998	2.42	1.6/1.4	0.070	0.082	−0.016
	Software classes	Directed	1 376	2 213	1.61	1.000	5.40	–	0.033	0.012	−0.119
	Electronic circuits	Undirected	24 097	53 248	4.34	1.000	11.05	3.0	0.010	0.030	−0.154
	Peer-to-peer network	Undirected	880	1 296	1.47	0.805	4.28	2.1	0.012	0.011	−0.366
Biological	Metabolic network	Undirected	765	3 686	9.64	0.996	2.56	2.2	0.090	0.67	−0.240
	Protein interactions	Undirected	2 115	2 240	2.12	0.689	6.80	2.4	0.072	0.071	−0.156
	Marine food web	Directed	134	598	4.46	1.000	2.05	–	0.16	0.23	−0.263
	Freshwater food web	Directed	92	997	10.84	1.000	1.90	–	0.20	0.087	−0.326
	Neural network	Directed	307	2 359	7.68	0.967	3.97	–	0.18	0.28	−0.226

Figure: The above table is from (Newman, 2010)

The properties measured in the previous table

- type of network
- directed or undirected
- total number of vertices n
- total number of edges m
- mean degree c
- fraction of vertices in the largest component S (or the largest weakly connected component in the case of a directed network);
- mean geodesic distance between connected vertex pairs ℓ
- exponent α of the degree distribution if the distribution follows a power law (or - if not; in/out-degree exponents are given for directed graphs);
- local clustering coefficient C:
- Average local clustering coefficient over all nodes
- the degree correlation coefficient r

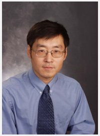
ER network vs real network

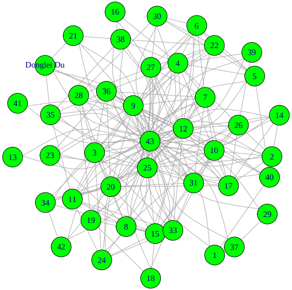
Characteristics	ER prediction	Real network
Density	$p \Longrightarrow$ Sparse	Sparse
Degree distribution	Poisson (or Normal)	Power-law
Clustering coefficient	$p \Longrightarrow$ Low	High
Average distance	Small world	Small world
Giant component	Yes	Yes
Community structure	No	Yes
Homophily	No	Yes

Case study: calculate the different measures for the Padgett Florentine families social network

```
rm(list=ls()) # clear memory
library(igraph) # load package igraph
load("padgett.RData") # read in the data
gb<-padgett$PADGB # The business network
#gm<-padgett$PADGM # the marriage network
##############################################################################
#Calculate the different measures for the Business network
##############################################################################
# calculate the degree and degree distribution
degree.distribution(gb)
degree(gb,loops = FALSE)
# calculate the path distribution:
shortest.paths(gb)
average.path.length(gb)
path.length.hist(gb) # $res is the histogram of distances,
  # $unconnected is the number of pairs for which the first vertex is not
  # reachable from the second.
# Calculate the clustering coefficient
transitivity(gb, type="local")# individual clustering
transitivity(gb, type="average") #average clustering
transitivity(gb)# overal clustering: the ratio of the triangles
   # and the connected triples in the graph.
```

Donglei Du's ego network on Facebook as of Sept 17, 2014





The size of the giant component Newman (2010)-Chapter 12

- $s = 1 - u$: the asymptotic ($n \to \infty$) fraction of vertices that are in the giant component S:

$$\boxed{s \approx 1 - e^{-\langle k \rangle s}} \tag{1}$$

- u: the probability that a randomly chosen vertex in the graph does not belong to the giant component S:

$$\boxed{u \approx e^{-\langle k \rangle (1-u)}}$$

- For a randomly chosen node i, $i \notin S$ iff it is not connected to S via any other $n-1$ nodes.
- For every other node $j \neq i$,
  - either: i is not connected to j with probability $1-p$;
  - or: i is connected to j but $j \notin S$ with probability pu.
- Therefore

$$
\begin{aligned}
u &= (1-p+up)^{n-1} = \left(1 - \frac{\langle k\rangle}{n-1}(1-u)\right)^{n-1} \\
&\Updownarrow \\
\ln u &= (n-1)\ln\left(1 - \frac{\langle k\rangle}{n-1}(1-u)\right) \underbrace{\approx}_{n\to\infty} -\langle k\rangle(1-u) \\
&\Updownarrow \\
u &= e^{-\langle k\rangle(1-u)}
\end{aligned}
$$

Percolation threshold

$$\boxed{\text{There is a giant component}}$$
$$\Updownarrow$$
$$\boxed{u < 1}$$
$$\Updownarrow$$
$$\boxed{s > 0}$$
$$\Updownarrow$$
$$\boxed{\langle k \rangle > 1}$$

Lambert W function

- We need the following concept to solve the equation (1).
- The following equation's solutions are called the Lambert W functions:

$$ye^y = x \Longleftrightarrow y = W(x) \text{ or } y = W_{-1}(x)$$

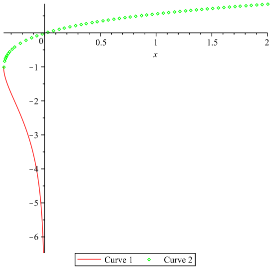


Figure: Lambert W function is defined only for $x \geq -e^{-1}$, and is double-valued for $x \in (-e^{-1}, 0)$. There are two solutions (1) $W(x)$ (green) refers to the principal branch satisfying $W(x) \geq -1$, and (2) $W_{-1}(x)$ (red) refers to the branch satisfying $w(x) < -1$.

Solution for (1) via Lambert W function

- The solution for (1) can be expressed via the Lambert W function:

$$\boxed{s = 1 - e^{-\langle k \rangle s}}$$

$$\Updownarrow$$

$$0 \underbrace{\geq}_{s \leq 1} \boxed{\langle k \rangle (s-1) e^{\langle k \rangle (s-1)} = -\langle k \rangle e^{-\langle k \rangle}} \underbrace{\geq}_{\langle k \rangle \geq 0} -e^{-1}$$

- The solution is:

$$\boxed{s = 1 + \frac{1}{\langle k \rangle} W(-\langle k \rangle e^{-\langle k \rangle}) > 0 \Longleftrightarrow \langle k \rangle > 1}$$

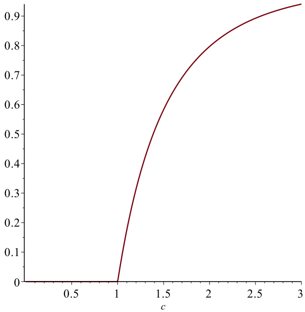


Figure: Size of the giant component s as a function of $c = \langle k \rangle$

Go Back

There is only one giant component!!!

- Suppose that there were two or more giant components in a random graph.
- Take any two giant components S_1 and S_2, with sizes $s_1 n$ and $s_2 n$ respectively $(s_1, s_2 \in [0,1])$.
- S_1 and S_2 are separate iff there is no edge connecting them together, which happens with probability q given by

$$q = (1-p)^{s_1 s_2 n^2} = \left(1 - \frac{c}{n-1}\right)^{s_1 s_2 n^2} = \Theta\left(e^{-c s_1 s_2 n}\right) \underbrace{\rightarrow}_{n \rightarrow \infty} 0$$

  - The number of distinct pairs of vertices (i,j), where $i \in S_1, j \in S_2$, is just $s_1 s_2 n^2$.
  - Each of these pairs is connected by an edge with probability p, or not with probability $1-p$.

The distribution of the sizes of the small components

- Let π_k be the probability that a randomly chosen vertex belongs to a small component of size exactly k vertices. Then

$$\sum_{k=0}^{\infty} \pi_k = 1 - s$$

- Claim: the potability distribution of the sizes of the small components in a random graph with mean degree c is given by

$$\pi_k = \frac{e^{-ck}(ck)^{k-1}}{k!}, k = 0, 1 \ldots.$$

Albert-Lszl Barabsi at TEDMED 2012

- http://www.youtube.com/watch?feature=player_detailpage&v=10oQMHadGos

References I

Bollobás, B., FULTON, W., KATOK, A., KIRWAN, F., and SARNAK, P. (2001). Cambridge studies in advanced mathematics. In *Random graphs*. Cambridge University Press New York.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Newman, M. (2010). *Networks: an introduction*. Oxford University Press.