A Short Introduction to Networks and Model Comparisons

Levi Lee
Advisor: Amy Wagaman
Amherst College

April 12, 2017

A Highly Connected World

- Networks are everywhere

History

- Travers & Milgram (1967)
- Letter correspondence between strangers in Nebraska and Massachusetts
- Overall, it took only around six people for the letter to be delivered
- "Six degrees of separation"

Social Networks

- Involves a lot of people
- Highly condensed groups
- Relatively short distances between people

Question

- Can we simulate this? If so, how?
- Conduct a simulation study utilizing different graph models

Basic Terminlogy

- A *graph*, denoted $G(V, E)$, consists of a set of *vertices* $i, j, k, ... \in V$ and a set of *edges* $\{i, j\}, \{i, k\}, \{j, k\}... \in E$.
- Our focus is on *simple*, *undirected*, and *connected* graphs.

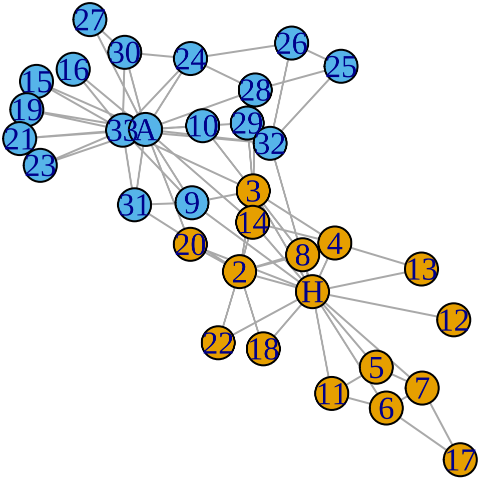
Other Representations of Graphs

- *Adjacency matrix*: denoted **A** is an N_V x N_V matrix where each element denotes the existence of edges between pairs where

$$A_{ij} = \begin{cases} 1 & \text{if } \{i, j\} \in E \\ 0 & \text{otherwise.} \end{cases}$$

- *Edge list*: two-column list of all the edges in a graph denoted by their corresponding vertices present

Example: Karate Club of Zachary (1977)



Network Statistics

- Analogous to statisitcs seen in elementary statistics
- Characterizes a given network

Transitivity/Clustering Coefficient

- Ratio of triangles to connected triples
- *Triangle*: three vertices connected by three edges
- *Connected triple*: three vertices connected by two edges

$$C = \frac{(\text{number of triangles}) \times 3}{\text{number of connected triples}}$$

Notions of Distance

Average path length: average of the shortest paths of all distinct pairs of vertices in the network

Diameter: longest of all the shortest paths between distinct pairs of vertices

Example: Karate Club of Zachary and Lazega's Law Firm

Network Statistic	Zachary's Karate Club
Transitivity	0.256
Average Path Length	2.408
Diameter	13

Centrality

- Measure of importance for each vertex in the graph
- Many different types of centralities exist

Degree Centrality

- Based on the number of edges are connected to a vertex
- Vertices with higher vertex degrees are considered to be more central to the network than those with lower vertex degrees

Closeness Centrality

- Measures how close a vertex is to other vertices based on the inverse of the total distance of the vertex from all others

$$c_{Cl}(i) = \frac{1}{\sum_{j \in V} d(i,j)}$$

- $dist(i,j)$ is the geodesic distance between the vertices $i, j \in V$

Betweenness Centrality

- Measures the extent to which a vertex is located between other pairs of vertices

$$c_B(i) = \sum_{g \neq h \neq i \in V} \frac{\sigma(g,h|i)}{\sigma(g,h)}$$

- $\sigma(g,h|i)$ is the total number of shortest paths between g and h that pass through i, and $\sigma(g,h) = \sum_{i \in V} \sigma(g,h|i)$

Eigenvector Centrality

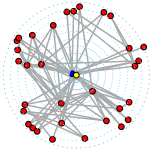
- Based on the idea of "status," "prestige," or "rank;" the more central the neighbors of a vertex are, the more central that vertex itself is

$$c_{Ei}(i) = \alpha \sum_{\{i,j\} \in E} c_{Ei}(u)$$

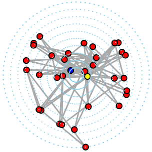
- $c_{Ei} = (c_{Ei}(1), ..., c_{Ei}(N_V))^T$ is the solution to the eigenvalue problem $\mathbf{A}\mathbf{c}_{Ei} = \alpha^{-1}\mathbf{c}_{Ei}$, where $\mathbf{A}$ is the adjacency matrix for network graph G.

Example: Karate Club of Zachary (1977)

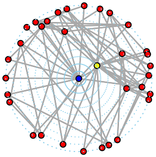
Degree



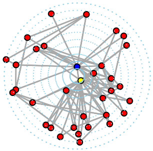
Closeness



Betweenness



Eigenvector



Example: Karate Club of Zachary

Network Statistic	Zachary's Karate Club
Avg. Degree	4.588
Avg. Closeness Cen.	0.005
Avg. Betweenness Cen.	26.194
Avg. Eigenvector Cen.	0.377

Graph Models

- A *graph model* takes in fixed parameters and generates a graph that vary in structure with each iteration
- Equivalently, it is a collection, or *ensemble* of graphs, denoted by

$$\{\mathbb{P}_\theta(G), G \in \mathcal{G} : \theta \in \Theta\}$$

- $\mathcal{G}$ is a collection or ensemble of possible graphs, P_θ is a *probability distribution* on the random graph G, and θ is a vector of parameters that describe the graphs that G can be, ranging over possible parameters in Θ

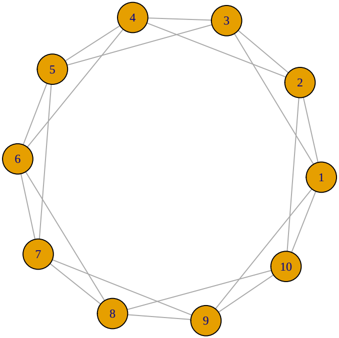
Erdős–Rényi Model (1959)

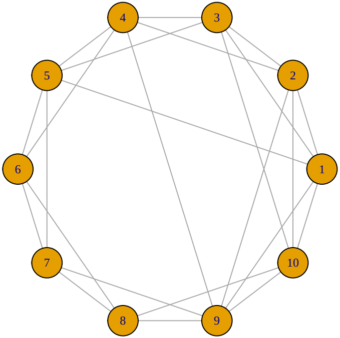
- Model with parameters: N_V, and N_E or p
- Model of the form $G(N_V, p)$ and $G(N_V, N_E)$

Properties of the Erdős–Rényi Model

- Short average path lenths
- Low clustering coefficient
- For the $G(N_V, p)$ model, a particular simple graph g with exactly N_V vertices has probability $P(G = g) = p^{N_E}(1-p)^{\left(\binom{N_V}{2} - N_E\right)}$

- Model with parameters: N_V, r, p





Properties of the Watts-Strogatz Model

- High clustering coefficient
- Small average path length

Exponential random graph models (ERGMs) I

- Exponential random graph models (ERGMs) are a class of models that can be used to generate probability distributions
- Flexible in design; we can decide our parameters
- Conduct goodness-of-fit tests for model assessment

Exponential random graph models (ERGMs) II

- The general form for an ERGM is as follows:

$$P_{\theta,\mathcal{G}}(\mathbf{G} = \mathbf{g}) = \frac{exp(\theta^T \mathbf{s(g)})}{\kappa(\theta, \mathcal{G})}, \mathbf{g} \in \mathcal{G}$$

- **Y** is the random variable representing a random graph and **g** is the particular adjacency matrix we observe. **s(g)** is the vector of model statistics for **g**, θ is the vector of coefficients for those statistics, and $\kappa(\theta, \mathcal{G})$ is the quantity in the numerator summed over all possible networks

Properties of ERGMs I

- Deriving the Erdős-Rényi Model from ERGMs
- Suppose we have a particular graph g and the only statistic we have is $L(G)$, the number of edges in g

$$P_{\theta,\mathcal{G}}(g) = \frac{exp(\theta_L L(g))}{\sum_{g' \in \mathcal{G}} exp(\theta_L L(g'))} = \frac{exp(\theta_L L(g))}{\kappa(\theta, \mathcal{G})}, g \in \mathcal{G}$$

Properties of ERGMs II

- Consider the probability distribution for a particular graph g with N_E edges again (from the Erdős-Rényi model). Using the fact that $N_E = L(g)$, taking the equation as a power of base e, we get the following:

$$
\begin{aligned}
P(g) &= p^{N_E}(1-p)^{\left(\binom{N_V}{2}-N_E\right)} \\
&= p^{L(g)}(1-p)^{\left(\frac{N_V(N_V-1)}{2}-L(g)\right)} \\
&= \left(\frac{p}{1-p}\right)^{L(g)}(1-p)^{\frac{N_V(N_V-1)}{2}} \\
&= exp\left(L(g)log\left(\frac{p}{1-p}\right) - \frac{N_V(N_V-1)}{2}log\left(\frac{p}{1-p}\right)\right) \\
&= exp(\theta_L L(g) - c) \\
&= \frac{exp(\theta_L L(g))}{exp(c)},
\end{aligned}
$$

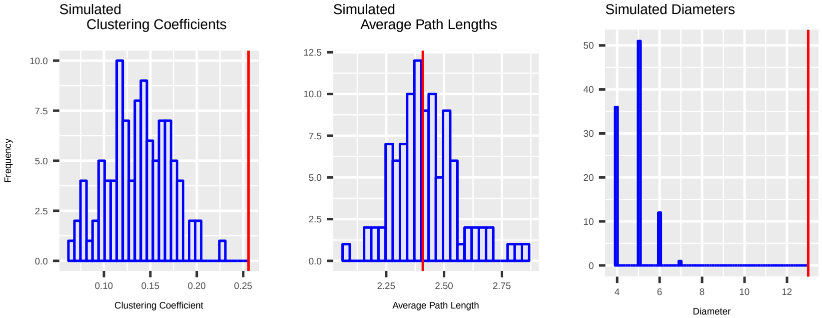
Data Set

- A component of the Facebook network
- 4039 vertices and 88234 edges
- Simple, connected, and undirected

Generating Random Graphs

- Create graphs of parable magnitude using certain parameters
- Simulate 1000 random graphs for and record network statatistics for each graph
- Create distribution of these values and/or see a table of averages
- Compare network statistics of the Facebook network with that of the graphs we generated

The Picture in Mind



Erdős–Rényi Model

- N_V and p
- The number of vertices is 4039
- Estimate the probability by taking the number of observed edges and dividing by the number of possible edges $\hat{p} = \frac{88234}{\binom{4039}{2}} = 0.011$
- For every possible edge among the 4039 vertices, determine if an edge will form based on the estimated probability

Watts-Strogatz Model

- N_V, r, p
- Will not use p
- Start with a lattice with 4039 vertices
- Randomly add $88234 - 4039 = 84195$ edges until we have approximately the same number as our observed network.
- Assign a number of edges to the vertices equal to the smallest degree observed in our Facebook network; 1
- Simplify our simulated graph to eliminate multi-edges and loop

Results for Erdős–Rényi and Watts-Strogatz Models

Network Statistic	Observed	Erdős-Rényi	Watts-Strogatz
Transitivity	0.617	0.0108 $\pm$ 0.0001	0.0107 $\pm$ 9.135e-05
Average Path Length	4.338	2.606 $\pm$ 0.002	2.6093 $\pm$ 0.0002
Diameter	17	3.96 $\pm$ 0.21	3.95 $\pm$ 0.22
Avg. Degree Cen.	43.691	43.69 $\pm$ 0.14	43.45 $\pm$ 0.01
Avg. Betweenness Cen.	2072.642	3242 $\pm$ 4	3249.2 $\pm$ 0.4
Avg. Closeness Cen.	8.881e-08	9.507e-05 $\pm$ 7.230e-08	9.494e-05 $\pm$ 7.319e-09
Avg. Eigenvector	0.040	0.620 $\pm$ 0.022	0.6235 $\pm$ 0.0227

ERGMs

- Four different ERGMs–labeled as ERGM 1a, ERGM 2a, ERGM 2b, and ERGM 3a
- ERGM 1a: one parameter: edges
- ERGM 2a: two parameters: edges and triangles
- ERGM 2b: two parameters: edges and k-stars (of size 3)
- ERGM 3a: three parameters: edges, triangles, and k-stars (of size 3)
- For each random graph, calculate the networks statistics of interest

Results for ERGMs I

Network Statistic	Observed	ERGM1a	ERGM2a
Transitivity	0.617	0.3696 ± 0.0012	0.4823 ± 0.0020
Average Path Length	4.338	2.885 ± 0.004	3.052 ± 0.0086
Diameter	17	5.216 ± 0.412	6.098 ± 0.035
Avg. Degree	43.691	43.66 ± 0.052	44.54 ± 0.03
Avg. Betweenness Cen.	2072.642	3805 ± 9	4140 ± 18
Avg. Closeness Cen.	8.881e-08	8.286e-05 $\pm$ 8.353e-06	6.008e-05 $\pm$ 1.514e-05
Avg. Eigenvector Cen.	0.040	0.0417 ± 0.0008	0.0410 $\pm$ 3.577e-05

Results for ERGMs II

Network Statistic	Observed	ERGM2b	ERGM3a
Transitivity	0.617	0.3787 ± 0.0013	0.4891 ± 0.0020
Average Path Length	4.338	2.851 ± 0.003	3.0562 ± 0.0079
Diameter	17	5.095 ± 0.293	6.161 ± 0.3677
Avg. Degree Cen.	43.691	44.06 ± 0.057	44.52 ± 0.034
Avg. Betweenness Cen.	2072.642	3736 ± 6	4148 ± 16
Avg. Closeness Cen.	8.881e-08	8.558e-05 $\pm$ 6.025e-06	5.875e-05 $\pm$ 1.5035e-05
Avg. Eigenvector Cen.	0.040	0.0392 ± 0.0002	0.0410 $\pm$ 3.451e-05

Conclusions and Future Work

- The models were bad, so now what?
- Choose other models
- Choose different network statistics
- Choose other data sets

Implications

- Understand the flow of information
- Better access to jobs through networking
- Better leads to resources in research
- Improving traffic
- Understanding biological systems

Sources I

Baumer, B., Kaplan, D., & Horton, N. (2017). Modern Data Science With R: With Digital Download. Taylor & Francis. Retrieved from https://books.google.com/books?id=Gv1nvgAACAAJ

Butts, C. T., & others. (2008). Social network analysis with sna. Journal of Statistical Software, 24 (6), 1–51.

Butts, C., Hunter, D., Handcock, M. S., Morris, M., Krivtisky, P. N., Almquist, Z.,. . . Bender de-Moll, S. (2015, June). Introduction to Exponential-family Random Graph (ERG or p*) modeling with ergm. Retrieved from https://statnet.org/trac/raw-attachment/wiki/Sunbelt2015/ergm_tutorial.pdf

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., Morris, & Martina. (2008). Ergm: A package to fit, simulate and diagnose exponential-family models for networks. Journal of Statistical Software, 24 (3), 1–29.

Sources II

Leskovec, J., Chakrabarti, D., Kleinberg, J., & Faloutsos, C. (2005). Realistic, Mathematically Tractable Graph Generation and Evolution, Using Kronecker Multiplication. In Knowledge Discovery in Databases: PKDD 2005 (pp. 133–145). Springer, Berlin, Heidelberg. Retrieved from
`http://link.springer.com/chapter/10.1007/11564126_17`

Jackson, M. O. (2013). Social and economic networks: Models and analysis. StanfordUniversity; Coursera Course Lecture.

Kolaczyk, E. D. (2009). Statistical Analysis of Network Data: Methods and Models. Springer Science & Business Media.

Kolaczyk, E. D., & Csárdi, G. (2014).Statistical Analysis of Network Data with R.Springer.

Lazega, E., & Pattison, P. E. (1999). Multiplexity, generalized exchange and cooperation in organizations: A case study. Social Networks, 21 (1), 67–90.

Sources III

Leskovec, J., & Krevl, A. (2014, June). SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data.

Leskovec, J., Chakrabarti, D., Kleinberg, J., & Faloutsos, C. (2005). Realistic, Mathematically Tractable Graph Generation and Evolution, Using Kronecker Multiplication. In Knowledge Discovery in Databases: PKDD 2005 (pp. 133–145). Springer, Berlin, Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/11564126_17

Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., & Ghahramani, Z. (2010).

Kronecker Graphs: An Approach to Modeling Networks. Journal of Machine Learning Research, 11 (Feb), 985–1042. Retrieved from http://www.jmlr.org/55papers/v11/leskovec10a.html

Sources IV

Newman, M. (2010). Networks: An Introduction. OUP Oxford

Travers, J., & Milgram, S. (1967). The small world problem. Phychology Today, 1, 61–67.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. Nature, 393 (6684), 440–442. http://doi.org/10.1038/30918

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 33 (4), 452–473. http://doi.org/10.1086/jar.33.4.3629752