Modelling the Facbook Social Network: The Memoryless GEO-P Graph Model

Pardis Noorzad

Department of Mathematics
Ryerson University

SOGMSC'14 — May 21, 2014

Outline

Online social networks (OSNs)

- Facebook (undirected network)
- Twitter, Instagram (directed networks)
- why study them?
- interesting for social scientists, marketers
  - look for online communities, influential actors, spread of information
  - design **network algorithms** for these studies

Why model online social networks networks?

- to test an algorithm, we generate data
- to study social networks we **generate graphs**
  - can be used to study network evolution over time
  - real-data at that scale might be unavailable, difficult to access

Random number generation

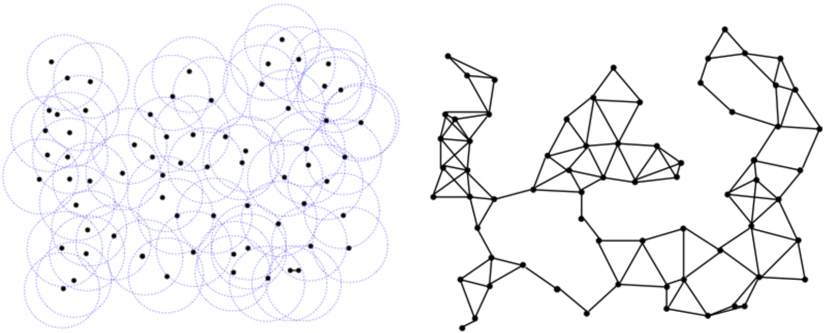
- common idea: generate random numbers
- similar ideas may be used generate a random binary matrix: each entry receives a 0 or 1 according to a fixed distribution
- **random matrices** correspond to **adjacency matrices** of directed graphs
- later we see that this is the Erdős-Rényi model for random graphs

Properties of OSNs

- **Large scale**. many nodes and many edges
- **Small world**. low distances between nodes, and high local clustering
- **Power law**. exhibit a degree distribution that is heavy tail

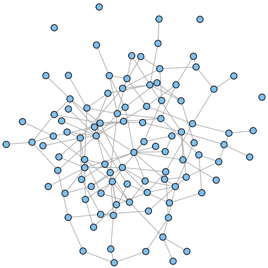
Background on random graphs

- definitive work of Erdős and Rényi (1959):
  the Erdős-Rényi or the binomial random graph
  - denoted $\mathcal{G}(n, p)$
    graph of order n
    edges added independently with probability p, $p \in (0, 1)$
- random **geometric** graph (Penrose, 2003)
  - denoted $\mathcal{G}(n, r)$
    n vertices u.a.r. on a unit hypercube
    an edge exists between two vertices if distance less than r

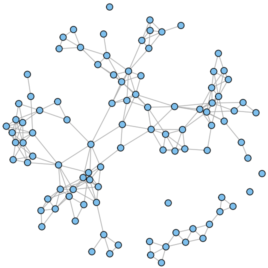


(a) Points placed uniformly at random. (b) Edges of a random geometric graph.

Figure : Random geometric graph in two dimensions (Diaz, 2008).



(a) Erdős-Rényi graph.



(b) Random geometric graph.

Figure : A comparison of the Erdős-Rényi and random geometric graphs.

MGEO-P(α, β, m, p)

Bonato, Gleich, Kim, Mitsche, Pralat, Tian, and Young (2014+)

"Memoryless", "Geometric" and "protean"

- $\alpha \in (0, 1)$ is the attachment strength,
- $\beta \in (0, 1 - \alpha)$ is the density parameter,
- $m \in \mathbb{N}$ is the dimension of the graph, and
- $p \in (0, 1]$ is the link probability.

MGEO-P model

Given an initial configuration of n vertices on a unit hypercube

1. fix a permutation σ on $\{1, \cdots, n\}$
   - where $\sigma(i)$ represents the rank of the ith oldest node
2. for each pair $i > j$, the edge $\{i, j\}$ is present iff
   - node i is in the ball of volume $\sigma(j)^{-\alpha}n^{-\beta}$

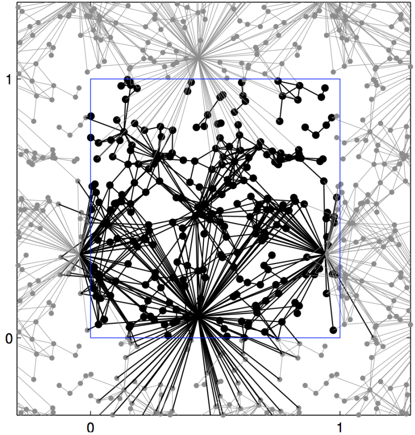
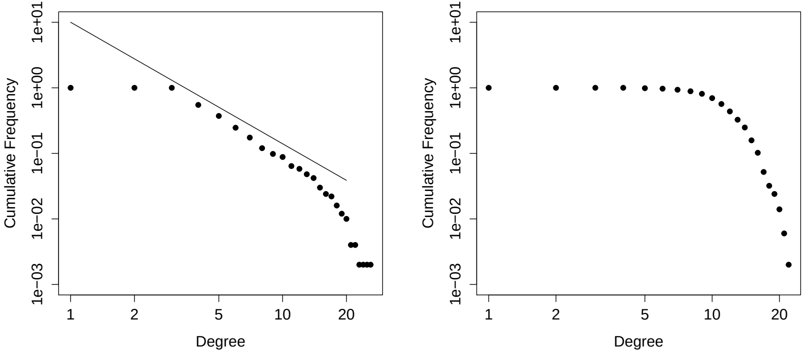


Figure : MGEO-P graph with $n = 250$ in two dimensions (Bonato, Gleich, Kim, Mitsche, Pralat, Tian, and Young, 2014+)

How good is your model?

- An important question: which model better fits the data?
- Isomorphism is too strong a similarity measure
  one obstacle is that the graphs are massive
- Instead consider graph properties: global and local
  - global: power law exponent
  - local: graphlet counts



(a) M-GEOP graph. (b) $\mathcal{G}(n, r)$ graph.

Figure : Log-log plot of the degree distributions for the M-GEOP and the $\mathcal{G}(n, r)$ models.

Local properties

- Many models share the power law property. Which one is a better fit for the FB100 data set?
- Consider graphlet counts for small order graphs (3 or 4) these capture local behaviour
- Wernicke's algorithm (Wernicke and Rasche, 2006)
    1. sample the graph
    2. search for graphlets in the sample and return a count

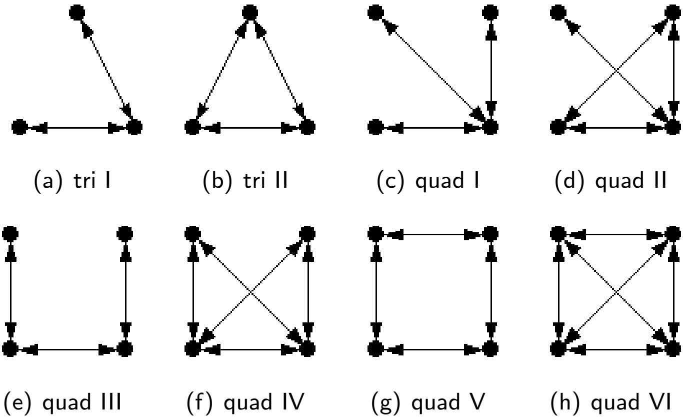


Figure : The graphlets of order 3 and 4.

Experiments

- **Data:** Facbook100 dataset
  - snapshot Facebook networks of 100 colleges from Sept. 2005
- **Algorithm:** For each graph in FB100
  1. extract its parameters:
     order, size, power law exponent, diameter
  2. Extract its graphlet count
  3. Generate samples from the M-GEOP model and samples from the $\mathcal{G}(n,r)$ model
  4. Percolate if necessary
  5. Extract graphlet counts for every sample
  6. Train classifier on graphlet count representation of the samples
  7. Use classifier to classify each FB graph
  8. Output whether $\mathcal{G}(n,r)$ or M-GEOP better fits the graph

Table : Statistics of the Facebook100 data set.

Name	Order	Size	PL exp	Eff Diam	α	β	m
Caltech36	769	16656	7.00	3.81	0.17	0.27	5
Reed98	962	18812	4.38	3.88	0.30	0.17	5
Harverford76	1446	59589	7.00	3.63	0.17	0.23	5
Simmons81	1518	32988	4.74	3.92	0.27	0.22	5
Swarthmore42	1659	61050	5.60	3.77	0.22	0.20	6
Amherst41	2235	90954	5.64	3.81	0.22	0.21	6
Bowdoin47	2252	84387	5.80	3.81	0.21	0.23	6
Hamilton46	2314	96394	4.63	3.79	0.28	0.15	6
...							

Table : Graphlet counts for the Facebook100 data set.

Name	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8
Caltech36	13	11	17	17	16	13	14	13
Reed98	13	11	17	18	17	14	14	12
Haverford76	15	13	19	20	18	16	16	14
Simmons81	14	12	18	18	17	14	15	13
Swarthmore42	15	13	19	20	19	16	16	14
Amherst41	15	13	20	20	19	16	17	15
Bowdoin47	15	13	20	20	19	16	17	15
Hamilton46	16	13	20	20	19	16	17	15
...								

Table : Classification results where A=mgeop, B=gnr

Name	Class	$P(A)$	$P(B)$
Caltech36	mgeop	11	17
Reed98	mgeop	11	17
Haverford76	mgeop	13	19
Simmons81	mgeop	12	18
Swarthmore42	mgeop	13	19
Amherst41	mgeop	13	20
Bowdoin47	mgeop	13	20
Hamilton46	mgeop	13	20
...			

Conclusion and future work

Result:

- samples belong to the MGEO-P class with high probability

Next steps:

- find 'best' classifier with cross-validation
- compare MGEO-P against other sophisticated models
- develop library to test stochastic models for any data set

References

Anthony Bonato. *A Course on the Web Graph*. The American Mathematical Society, 2008.

Anthony Bonato, Jeannette Janssen, and Pawel Pralat. The geometric protean model for on-line social networks. In *WAW*, pages 110–121, 2010.

Anthony Bonato, Jeannette Janssen, and Pawel Pralat. Geometric protean graphs. *Internet Mathematics*, 8(1-2):2–28, 2012.

Anthony Bonato, David Gleich, Myunghwan Kim, Dieter Mitsche, Pawel Pralat, Amanda Tian, and Stephen Young. Dimensionality matching of social networks using motifs and eigenvalues. 2014+.

Josep Diaz. Random Geometric Graphs. http://www.lsi.upc.edu/~diaz/RGG_HK.pdf, 2008. [Online; accessed March 2013].

Paul Erdős and Alfréd Rényi. On random graphs. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

Mathew Penrose. *Random Geometric Graphs (Oxford Studies in Probability)*. Oxford University Press, USA, 2003.

Sebastian Wernicke and Florian Rasche. Fanmod: A tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, May 2006.

Thank you!

Special thanks:

- Prof. Anthony Bonato
- SOGMSC'14 organizers
- Fields Institute