CMSC 828E: Graph Generators; Properties; Models

Amol Deshpande

University of Maryland, College Park

November 3, 2010

Overview

- Several related questions
  - Measuring and understanding graph properties
  - Generating graphs that match observed behavior
    - The model must provide intuitions
- Why study "graph generators"?
  - Insights into global structure; Null models
  - Simulations; Extrapolations; Sampling
  - Anonymization

Some key network properties

- Node degreees follow power law distributions
  - i.e., if N_d = number of nodes with degree d, then
  - $N_d \propto d^{-\gamma}, \gamma > 0$ is called *power law exponent*
  - These distributions have *heavy tails* (as opposed to Gaussian)
  - Also see: scale-free or scale invariance
- Aside: Power Laws
  - Plot the quantities (e.g., d vs N_d) on a log-log scale
  - See if it is a straight line
  - Careful: Can be quite misleading
    - See: Power-Law Distributions in Empirical Data; Clauset et al.; 2009
    - Too often we see power laws when there isn't one
    - Or other models fit equally well

Some key network properties

- Small Diameter
  - Diameter = maximum distance between any two nodes
    - Sensitive to outliers
  - Effective diameter: 90% of the pairs of nodes are within that distance
  - Effective diameter very small for most real networks
    - *6 degrees of separation*
    - Also see: small world phenomenon
- Shrinking Diameter
  - Effective diameter shrinks over time
  - Somewhat counterintuitive

Some key network properties

- Densification laws
  - Average degree of a node in the network increases over time
- Large clustering coefficients
  - Clustering coefficient = $\frac{3 \times number\ \ of\ \ triangles}{number\ \ of\ \ connected\ \ triples}$
  - (u, v, w) is a connected triple if (u, v) and (v, w) are edges
  - Why ?
- Self-similarity/Fractal behavior
- See this paper for many other power laws and empirical properties
  - RTG: A Recursive Realistic Graph Generator using Random Typing; ECML/PKDD 2009

Model: Random Graphs

- Erdos-Renyi model: each edge with equal probability
  - Does not match observed behavior, e.g., power law degree distribution
- Configuration model
  - Choose a power law degree distribution first
  - Generate a graph that satisfies the degree distribution
    - Attach "half-edges" to each node as required; Randomly connect half-edges
  - Generates every possible topology with the given degree sequence with equal probability
- No transitivity
  - i.e., these graphs have low clustering coefficients
- Exponential Random Graphs solve some of these problems in theory, but are generally intractable

Small World Model; Watts and Strogatz

- Focuses on the "high transitivity" aspect
- Start with a low-dimensional lattice of some form
  - Most studied: A ring with edges between all nodes within distance *k*
  - For each edge, "rewire" (move one endpoint) randomly with probability *p*
- Variation: keep the initial lattice the same, but add random edges

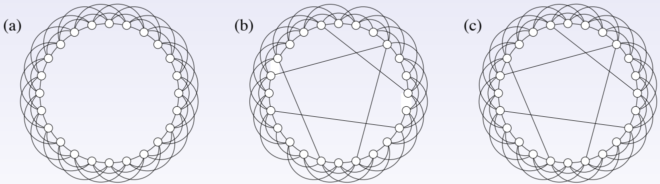


FIG. 11 (a) A one-dimensional lattice with connections between all vertex pairs separated by k or fewer lattice spacing, with $k = 3$ in this case. (b) The small-world model [412, 416] is created by choosing at random a fraction p of the edges in the graph and moving one end of each to a new location, also chosen uniformly at random. (c) A slight variation on the model [289, 324] in which shortcuts are added randomly between vertices, but no edges are removed from the underlying one-dimensional lattice.

Small World Model; Watts and Strogatz

- Focuses on the "high transitivity" aspect
- Start with a low-dimensional lattice of some form
  - Most studied: A ring with edges between all nodes within distance k
  - For each edge, "rewire" (move one endpoint) randomly with probability p
- If $p = 0$, then high clustering coefficient (for the ring, about 3/4)
  - But no "small world" effect – diameter is too high
- If $p = 1$, then almost random graph
  - Low diameter, but low clustering coefficient
- In between values of p create realistic networks
  - Clustering coefficient: $\frac{3(k-1)}{2(2k-1)}(1-p)^3$
  - Degree distributions not power law
  - Path lengths low

Model: Preferntial Attachment Model

- Barabasi and Albert; Science 1999 (some prior, forgotten, work)
- Popular and leads to power law degree distributions
- Prob. a new node attaches to an existing node $\propto$ degree of the node
- Results in a network with power law distribution
  - With exponent 2.8-3.0
- Many variations since (forest fire model can be seen as a variation on this)

Model: Duplicate Attachment Models

- Kumar et al.; FOCS 2000
- Originally developed for Web graphs, but also work well for Biological graphs
- A new node copies the neighbors of an existing node and attaches to that node
- Leads to communities??
    - Saw this mentioned, but did not find a proper analysis

Model: Community Guided Attachment

- Nodes are arranged in a community hierarchy
- For two nodes u and v, let $h(u, v)$ denote the height of the least common ancestor
- u and v are randomly connected with probability $f(h(u, v))$
  - $f(h) = c^{-h}$
- Leads to densification law
- Can be extended to a *dynamic* setting
  - Where we need to define how a new node enters the system
- Dynamic model also leads to heavy-tailed distribution of in-degrees

Model: Forest Fire

- The previous model did not capture:
  - Heavy-tailed out-degree distributions
  - Shrinking effective diameters
- The Basic Model:
  - Start with a single node graph
  - A new node v first chooses an existing node w uniformly at random and connects to it
  - Two random numbers are generated x and y
  - x outlinks and y inlinks of w are selected
  - Outlinks are formed to all those nodes
  - Process repeated

Model: Forest First

- Why properties arise?
  - Heavy-tailed in-degrees: highly-linked nodes can be easily reached by v
  - Communities: v starts by copying
  - Heavy-tailed out-degrees: reasonable chance of high number of outlinks
  - Densification power law: high numbers of links generated naturally
  - Shrinking diameter??
    - Not obvious but supported by analysis

Model: Kronecker Graphs

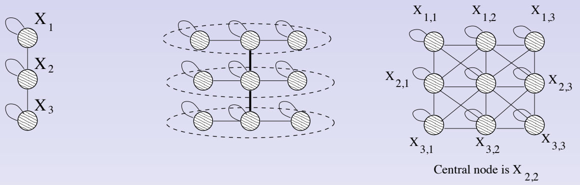
- Leskovec et al.; JMLR 2010
- A systematic and principled approach based on the notion of Kronecker graphs
- Basic Idea:
  - Start with a "seed" graph – usually a very small graph
  - Multiply with itself to generate large graphs
  - Kronecker multiplication:
    - given two matrices A and B of sizes $n \times m$ and $n' \times m'$

$$\mathbf{C} = \mathbf{A} \otimes \mathbf{B} \doteq \begin{pmatrix} a_{1,1}\mathbf{B} & a_{1,2}\mathbf{B} & \dots & a_{1,m}\mathbf{B} \\ a_{2,1}\mathbf{B} & a_{2,2}\mathbf{B} & \dots & a_{2,m}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1}\mathbf{B} & a_{n,2}\mathbf{B} & \dots & a_{n,m}\mathbf{B} \end{pmatrix}.$$

    - The result is a $nn' \times mm'$ matrix
  - For graphs, apply to "adjacency matrix"

Model: Kronecker Graphs

- Example



(a) Graph K_1 (b) Intermediate stage (c) Graph $K_2 = K_1 \otimes K_1$

1	1	0
1	1	1
0	1	1

(d) Adjacency matrix of K_1

K_1	K_1	0
K_1	K_1	K_1
0	K_1	K_1

(e) Adjacency matrix of $K_2 = K_1 \otimes K_1$

Figure 1: *Example of Kronecker multiplication:* Top: a “3-chain” initiator graph and its Kronecker product with itself. Each of the X_i nodes gets expanded into 3 nodes, which are then linked using Observation 1. Bottom row: the corresponding adjacency matrices. See Figure 2 for adjacency matrices of K_3 and K_4.

Model: Kronecker Graphs

- The generated graphs can be analytically shown to follow the properties like power law distributions etc.
- The discrete version not sufficiently powerful
    - Instead, use a small, say 2 by 2 matrix, with real numbers
    - Multiply with itself as many times as required
    - Treat the "numbers" in the result matrix as probabilities to generate a sample graph
- The above is called "Stochastic Kronecker Graphs"

Model: Kronecker Graphs

- Interesting question: Given a real graph, can I find the "seed graph" for it?
  - Can phrase a maximum likelihood problem
  - Not easy, but the authors show how to do this
- Turns out 2 by 2 seed matrices do very well
  - In other words, with four parameters we can capture the real networks

Structure and Evolution of Online Social Networks

- Kumar et al.; KDD 2006
- Studied two networks in depth
  - Flickr and Yahoo! 360
  - Have entire lifetime, with precise timestamps
- Results
  - Density: rapid growth, decline, then steady growth
- Divide the users into three groups
  - Singletons – signed up, but zero degree
  - Giant Component
  - Middle region

Structure and Evolution of Online Social Networks

- Kumar et al.; KDD 2006
- Middle region consists largely of "stars"
  - One or two center individuals with everyone else connected to them
- Over time, small communities and singletons blend into the giant component
- Leads to a graph evolution model where:
  - A new node is given a role when it comes into the system
  - A bunch of edges also arrive, possibly assigned between existing nodes

Community Structure in Large Networks

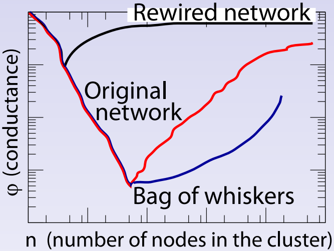
- Leskovec et al.; WWW 2008
- Defined the notion of "network community profile plot"
  - For each k, find the best community of size k
  - Defined to be the group of k nodes with the lowest conductance to the rest of the graph
- Observed that the curve bottoms out at around 100
  - Consistent with social observations that max community sizes are around 100-150
- Whiskers: Group of nodes connected to the rest of the graph by very few edges
  - Many 1-whiskers in social networks
- Leads to a core+whiskers model
  - None of the models explain this
  - Forest Fire model does a pretty good job

Community Structure in Large Networks

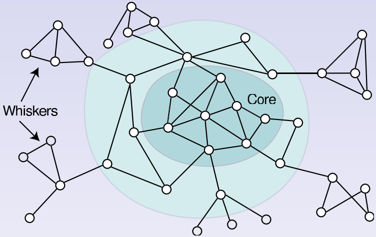
- Leskovec et al.; WWW 2008
- Many other observations
  - For small networks, we don't see the above effect
    - Communities are too small to blend into the core
  - The core itself is still very sparse
    - Core itself has a core+periphery structure (self-similar)
    - Core tends to be like an "expander"

Community Structure in Large Networks

- Leskovec et al.; WWW 2008



(a) Typical NCP plot



(b) Caricature of network structure

Figure 2: (a) Typical network community profile plot for a large social or information network: networks have better and better communities up to a size scale of ≈ 100 nodes, and after that size scale communities "blend-in" with the rest of the network (red curve). However, real networks still have more structure than their randomized (conditioned on the same degree distribution) counterparts (black curve). Even more surprisingly, if one allows for disconnected communities (blue curve), the community quality scores often get even better (even though such communities have no intuitive meaning). (b) Network structure for a large social or information network, as suggested by our empirical evaluations. See the text for more information on the "core" and "whiskers," and note that the core in our real-world networks is actually extremely sparse.