# CSCE 619-600
# Networks and Distributed Processing
# Spring 2017

## Random Graphs V

Dmitri Loguinov
Texas A&M University

April 4, 2017

# Agenda

- Static Models
  - Waxman
  - PLRG
  - GED
- Evolving models
  - Barabasi-Albert (BA)
  - GLP
  - HOT

# Static Topology Models

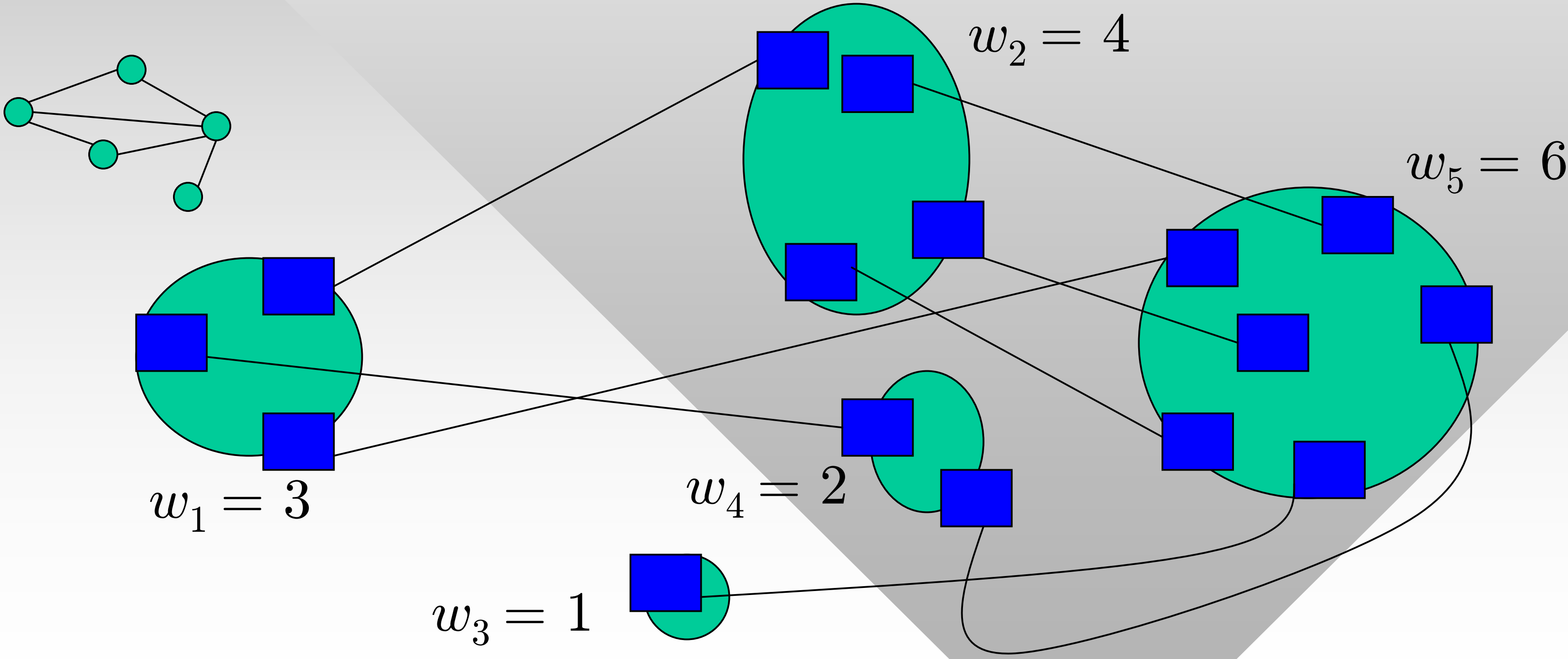
- Traditional networking models (e.g., Waxman 1988) have an exponentially-decaying tail of degree distribution
- Place $N$ points on a 2D plane
  - Create an edge between $(x,y)$ with probability $p_{xy} = 2^{-d(x,y)}$, where $d(x,y)$ is the Euclidian distance from $x$ to $y$
  - This approach discourages long-distance links and models the extra cost of leasing bandwidth over longer haul
- Additional modifications include $p_{xy}$ being proportional to $1/d(x,y)$ and other functions
  - In any case, degree distribution of Waxman topologies is not power-law, but exponential

# Static Topology Models 2

- PLRG (Power-Law Random Graph)
  - Aiello *et al*. in STOC 2000
- Generate a sequence of node weights $w_i$ according to the desired Pareto distribution
  - Each node $i$ gets some weight $w_i$
  - The total number of nodes is $N$
- Create a new graph with $w_i$ copies of each node $i$
- Then randomly generate edges between these virtual nodes with a uniform probability
  - Every time a pair of virtual nodes is connected, they are removed from the graph

# Static Topology Models 3

- Finally collapse the constructed graph
  - Distribution of resulting degrees $d_i$ will be similar to that of $w_i$
- Example:



# Static Topology Models 4

- Problems with PLRG
  - Does not guarantee connected graphs
  - Static model (i.e., no growth or evolution of the graph)
  - Produces self-loops and duplicate links whose removal may skew the degree distribution
- The original authors proved that there was always a giant connected component in PLRG
  - Usually contained $80$-$90\%$ of $N$
- They also showed that the remaining isolated subgraphs were $O(1)$ in size with high probability

# Static Topology Models 5

- Another approach is to extend $G(n,p)$ to achieve a heavy-tailed degree distribution
  - This model is called GED (Given Expected Degree)
- We first generate $n$ weights $w_i$ from the desired Pareto distribution and randomly assign them to nodes
  - Then insert each edge $(i,j)$ into the graph with probability

$$p_{ij} = \min\left(\frac{w_i w_j}{\sum_{k=1}^{n} w_k}, 1\right)$$

- The $\min$ function is needed when the sequence is so heavy-tailed that $w_i w_j$ is larger than the weight sum

# Static Topology Models 6

- Easy to notice that the expected degree of the constructed graph has the same distribution as $\{w_i\}$:

$$E[d_i] = \sum_{j=1}^{n} p_{ij} = w_i \frac{\sum_{j=1}^{n} w_j}{\sum_{k=1}^{n} w_k} = w_i$$

- Same drawbacks as in PLRG
  - Disconnected graphs are possible
- GED similarly does not explain the *growth* of the Internet over time
  - The Internet did not happen all at once, but rather evolved into its current shape

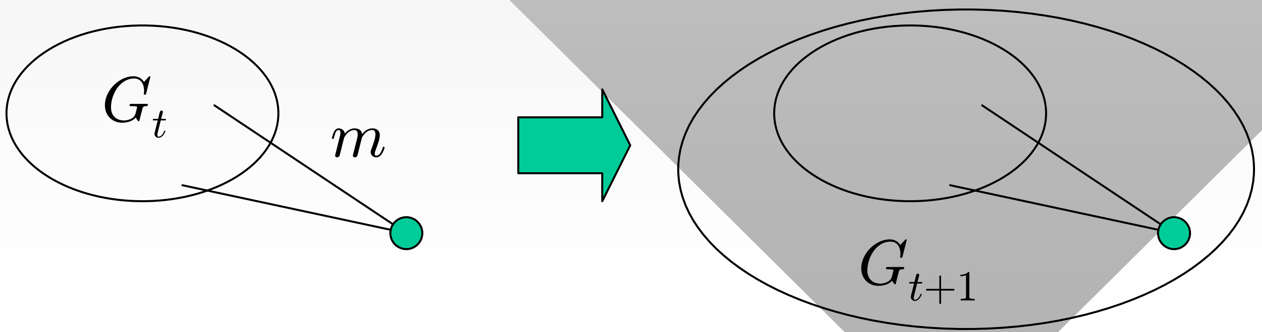
# Static Topology Models 7

- Thus, it is beneficial to examine evolving models that can achieve the Pareto degree distribution
  - Need a mechanisms allows us to built a series of graphs $G_t$ as a function of time $t$, where each $G_t$ has a heavy-tailed degree distribution and is obtained from $G_{t-1}$ by adding one node and some edges

- This is similar to the graph process discussed earlier for $G(n, M)$; however, the goal here is to keep the degree distribution *time-invariant*
  - Such graphs are known as scale-free since their degree distribution does not depend on time or scale at which they are examined

# BA Model

- The BA model starts with $m_0$ nodes fully connected
  - This is a complete graph representing the initial system where everybody "knows" everybody else
- At each step $t > m_0$, a new node joins the system and attaches itself to existing nodes through $m$ edges (there are $t$–$1$ existing nodes by time $t$)
  - Parameter $m$ is usually $1$–$3$ edges



# BA Model 2

- Since a new node is added at each step, the total number of steps is $n - m_0$
- Note: there is a small possibility that we may add duplicate edges when $m \geq 2$
  - This probably tends to zero as graph size increases and is generally ignored in the analysis
- The final question is how to attach the new edges
- Suppose the degree of node $i$ at time $t$ is $D_i(t)$
  - Then the new node that joins at time $t$ connects to $i$ with probability:

$$p_i(t) = \frac{D_i(t)}{\sum_{k=1}^{t-1} D_k(t)}$$

# BA Model 3

- This is the so-called preferential attachment
  - The new nodes perceive large-degree nodes that are already present in the system as being more “attractive” than their low-degree counterparts
- For social networks, this theory means that “popular” people are more attractive
- For the Internet, this theory says that larger ISPs attract connections from new ISPs with a higher probability
  - Large degree can imply better network design, longer presence on the market, better routing conditions, better connectivity to other ISPs, etc.

# BA Analysis

- Assume that we label nodes in the order in which they joined the graph
  - Thus, node $i$ joins at time $t_i = i$
  - The first $m_0$ nodes are an exception and their labeling is not relevant here
- Question: if at large time $t$ we uniformly randomly select a node $x$ from the system, what is the probability that $t_x$ is less than some constant $a \leq t$?

$$P(t_x \leq a) = \frac{a}{t}$$

  - Since $t_i$ is discrete and uniform in $[1, t]$

# BA Analysis 2

- What is the probability that a new node selects $i$ as a peering point after adding $m$ edges?
- This is approximately:

$$1 - \left(1 - p_i(t)\right)^m \approx mp_i(t)$$

- Next, we can write a simple stochastic difference equation on $D_i(t)$ for $i < t$:

$$D_i(t+1) - D_i(t) = \begin{cases} 1 & \text{w.p. } mp_i(t) \\ 0 & \text{w.p. } 1 - mp_i(t) \end{cases}$$

- Considering $n \to \infty$, in expectation this becomes:

$$E[D_i(t+1) - D_i(t)] = mE[p_i(t)] \approx \frac{mE[D_i(t)]}{2mt}$$

# BA Analysis 3

- Define $d_i(t) = E[D_i(t)]$ and write a continuous version:

$$\frac{\partial d_i(t)}{\partial t} = \frac{d_i(t)}{2t}$$

- This technique (transition from discrete to continuous equations) is often called *continuum theory* in physics and fluid model in networking
- Thus, we can view $d_i(t)/2t$ as the *instantaneous rate* of increase in degree at node $i$ at time $t$
  - Clearly, the larger $d_i(t)$, the faster $i$'s degree grows
  - On average, older nodes have larger degree at any $t$

# BA Analysis 4

- Solving $dy/dt = y/(2t)$, we obtain $y = C\sqrt{t}$
- Since the initial degree is $m$, we get the final version:

$$d_i(t) = m\sqrt{\frac{t}{t_i}}$$

- The degree of node $i$ grows as a square root of time from the moment it joined the graph
- We are finally ready to compute the distribution of $d_i(n)$ at the end of the construction process:

$$P(d_i(n) < x) = P(m\sqrt{n/t_i} < x)$$

# BA Analysis 5

- Expanding the last probability:

$$P(d_i(n) < x) = P(t_i > m^2 n/x^2)$$

  - which can be further written as:

$$P(d_i(n) < x) = 1 - P(t_i < m^2 n/x^2) = 1 - P(t_i < a)$$

- As we already know $P(t_i < a) = a/n$, we can use it in the above:

$$P(d_i(n) < x) = 1 - \frac{m^2 n}{x^2 n} = 1 - \frac{m^2}{x^2}$$

- This is a Pareto distribution with $\alpha = 2$ and $\beta = m$:

$$P(X < x) = 1 - (x/\beta)^{-2}$$

# BA Analysis 6

- Limitation of BA
  - The model always builds Pareto graphs with $\alpha = 2$
  - Ideally, we would like it to achieve arbitrary $\alpha > 1$
- Another problem with BA is that its clustering is virtually zero ($\gamma(G) \sim \log^2(n)/n$)
  - Its local structure is similar to that in $G(n,p)$ and thus very different from that of real graphs
- We can overcome both issues with using a slightly different preference function

# GLP

- GLP (Generalized Linear Preference) improves over BA in both shape $\alpha$ and clustering
- The idea is to change the attachment function to:

$$p_i(t) = \frac{D_i(t) - \lambda}{\sum_{k=1}^{t-1}(D_k(t) - \lambda)}, \lambda \in (-\infty, 1]$$

- Shape of the degree distribution is $\alpha = 2 - \lambda \in [1, \infty)$
- In addition, clustering is much higher than in BA
  - No closed form expression, but was experimentally observed to be as high as 0.35 for certain $n$
- Diameter is close to optimal as well

# HOT

- HOT (Highly Optimized Tradeoffs): alternative model
  - Applies to networks created as a result of some optimization
  - Optimization for ISPs may be the cost of links, available bandwidth, reachability, etc.

- Model description
  - Place $n$ points in a geographical (2D) area
  - Denote by $d_{xy}$ the Euclidean distance from $x$ to $y$ and by $h_{xy}$ the hop distance in the current graph from $x$ to $y$

- Define $h_j$ to be the average hop distance from some existing node $j$ to all other nodes in the graph:

$$h_j = \frac{1}{n} \sum_{i=0}^{n} h_{ji}$$

# HOT 2

- A new node $i$ attaches to node $k$ that minimizes the following:

$$k = \arg\min_{j<i}\{\theta d_{ij} + h_j\}$$

  - Where $\theta$ is a weight that balances the importance of "centrality" in the graph and link cost to the ISP
  - Minimization takes place over all nodes $j < i$ that joined before $i$

- An extension is possible with $m$ links added to each joining node

- HOT builds Pareto-degree graphs with $\alpha = 2$ and very high clustering (approx 0.5)