Outline

1. Introduction
   - Preferential Attachment (PA)
2. Common Neighbors Model (CN)
   - Degree distribution
   - Community structure
Preferential Attachment

- Users prefer to connect to nodes of high degree
Preferential Attachment

- Users prefer to connect to nodes of high degree
- Results in heavy-tailed degree distribution
Issues with Preferential Attachment

The LinkedIn graph

1. does NOT have a power law degree distribution
2. has “community structure”
Log-log plots of degree distribution

PA tail probability

Data tail probability
Issues with Preferential Attachmment

The LinkedIn graph

1. does NOT have a power law degree distribution
2. has “community structure”
What is “community structure”? 

- Strong community structure
- More edges within community than between communities
What is “community structure”?

- Preferential attachment
- One central hub around high-degree node
Common Neighbors Model
Common Neighbors Model

Users prefer to connect to nodes with whom they share many mutual friends.
Common Neighbors Model

Users prefer to connect to nodes with whom they share many mutual friends.
Common Neighbors Model

Sequence of graphs \((G_t)_{t \geq 0}\).

- Given graph \(G_t\) with \(n(t)\) nodes and \(m(t)\) edges
Common Neighbors Model

Sequence of graphs \((G_t)_{t \geq 0}\).

- Given graph \(G_t\) with \(n(t)\) nodes and \(m(t)\) edges
- At time \(t + 1\), a new node \(v\) arrives with probability \(\alpha\)
  - If no new arrival, select \(v\) uniformly among existing nodes

\(\Gamma(v)(t)\) is the neighborhood of \(v\) at time \(t\)

\(K_{vw}(t) = |\Gamma(v)(t) \cap \Gamma(w)(t)|\)

\[P(\text{select } w | \text{sender } = v) = K_{vw}(t) + \delta \sum u K_{vu}(t) + \delta n(t)\]

Form directed edge \((v, w)\).
Common Neighbors Model

Sequence of graphs \((G_t)_{t \geq 0}\).

- Given graph \(G_t\) with \(n(t)\) nodes and \(m(t)\) edges
- At time \(t + 1\), a new node \(v\) arrives with probability \(\alpha\)
  - If no new arrival, select \(v\) uniformly among existing nodes
- Select receiving node \(w\) with probability proportional to number of common neighbors between \(v\) and \(w\)
  - \(\Gamma_v(t)\) is the neighborhood of \(v\) at time \(t\)
  - \(K_{vw}(t) = |\Gamma_v(t) \cap \Gamma_w(t)|\)

\[
P(\text{select } w \mid \text{sender } = v) = \frac{K_{vw}(t) + \delta}{\sum_u K_{vu}(t) + \delta n(t)}
\]
Common Neighbors Model

Sequence of graphs \((G_t)_{t \geq 0}\).

- Given graph \(G_t\) with \(n(t)\) nodes and \(m(t)\) edges
- At time \(t + 1\), a new node \(v\) arrives with probability \(\alpha\)
  - If no new arrival, select \(v\) uniformly among existing nodes
- Select receiving node \(w\) with probability proportional to number of common neighbors between \(v\) and \(w\)
  - \(\Gamma_v(t)\) is the neighborhood of \(v\) at time \(t\)
  - \(K_{vw}(t) = |\Gamma_v(t) \cap \Gamma_w(t)|\)

\[
P(\text{select } w \mid \text{sender } = v) = \frac{K_{vw}(t) + \delta}{\sum_u K_{vu}(t) + \delta n(t)}
\]

- Form directed edge \((v, w)\).
Common Neighbors Model

What does $K_{vw}(t)$ look like?
Common Neighbors Model

What does $K_{vw}(t)$ look like?
Common Neighbors Model

What does $K_{vw}(t)$ look like?

Hard to analyze - feedback
Common Neighbor Process

- Want to model evolution of $K_{ij}(t)$ on its own.
- Start at $\tilde{K}_{ij}(0) = 0$ for all pairs $i, j$. 
Common Neighbor Process

- Want to model evolution of $K_{ij}(t)$ on its own.
- Start at $\tilde{K}_{ij}(0) = 0$ for all pairs $i, j$.
- Given $(\tilde{K}_{ij}(t))_{i,j \geq 0}$, at $t + 1$,
  - Select $i$ uniformly from existing nodes
Common Neighbor Process

- Want to model evolution of $K_{ij}(t)$ on its own.
- Start at $\tilde{K}_{ij}(0) = 0$ for all pairs $i, j$.
- Given $(\tilde{K}_{ij}(t))_{i,j \geq 0}$, at $t + 1$,
  - Select $i$ uniformly from existing nodes
  - Choose $\eta = c(n(t))^\theta$ nodes, $j_1, j_2, \ldots, j_\eta$, preferentially with $\tilde{K}_{ij_\ell}(t)$, and increase
    $$\tilde{K}_{ij_\ell}(t + 1) = \tilde{K}_{ij_\ell}(t) + 1.$$
Common Neighbor Process

Let

\[ N_i(t) = \sum_j \tilde{K}_{ij}(t) \]

What is the distribution of \( N_i(t) \)?
Common Neighbor Process

Theorem

Let $N_i(t) = \sum_j \tilde{K}_{ij}(t)$. Then there exists a random variable $Z_i$ such that

$$\frac{N_i(t)}{t^\theta} \to Z_i$$

in probability, where $Z_i$ has characteristic function

$$\phi_{Z}(z) = \exp \left\{ \frac{1 - \alpha}{\alpha \theta} \int_0^{\alpha \theta} \frac{1}{t} (e^{itz} - 1) \, dt \right\}.$$
Common Neighbor Process

Theoretical params: theta=1.4, alpha=.4

Empirical cdf of data
Theoretical cdf
Common Neighbor Process

Result

• The “total common neighbors” $N_i(t)$ converges when scaled by $t^\theta$.

In progress/Future

• Limiting distribution for $\tilde{K}_{ij}(t)$.
• Use these distributions to analyze degree distribution of the graph.
Community Structure

- How to quantify “strong community structure”
- Compare community structure of CN and PA.
Community Structure CN vs. PA

CN, 200 nodes, 500 edges, attraction=high

PA, 200 nodes, 500 edges, attraction=high
Modularity

Definition
Given a graph partitioned into $c$ communities, the modularity is

$$Q = \sum_{i=1}^{c} (e_{ii} - a_i^2)$$

where $e_{ii}$ is the fraction of edges with both end vertices in community $i$, and $a_i$ is the fraction of ends of edges with vertices in community $i$. 

Community Detection

• Community detection algorithms aim to assign nodes to communities in a way that is reasonable.

• Some algorithms maximize modularity: Fast-greedy (FG), Largest-eigenvector (LE).

• But there are other methods as well: Edge-betweenness (EB), Walktrap (WC).
Averages of modularity over 100 trials ($\alpha = .2, \delta = .5$)

<table>
<thead>
<tr>
<th>Graph</th>
<th>EB</th>
<th>FG</th>
<th>LE</th>
<th>WC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN 500</td>
<td>.450</td>
<td>.472</td>
<td>.423</td>
<td>.401</td>
</tr>
<tr>
<td>PA 500</td>
<td>.276</td>
<td>.379</td>
<td>.333</td>
<td>.251</td>
</tr>
<tr>
<td>CN 1000</td>
<td>.310</td>
<td>.402</td>
<td>.350</td>
<td>.301</td>
</tr>
<tr>
<td>PA 1000</td>
<td>.103</td>
<td>.328</td>
<td>.279</td>
<td>.190</td>
</tr>
<tr>
<td>CN 5000</td>
<td>.145</td>
<td>.320</td>
<td></td>
<td>.176</td>
</tr>
<tr>
<td>PA 5000</td>
<td>.039</td>
<td>.277</td>
<td></td>
<td>.120</td>
</tr>
</tbody>
</table>
Conclusion

1. PA mode lacks characteristics of LinkedIn network:
   - Power-law degree distribution
   - Lack of community structure

2. Common Neighbors Model
   - Limiting distribution of $N_i(t)$ in the common neighbors process
   - Better community structure than PA
Edge Acceptance/Rejection

Node $v$ sends an invitation to a node $w$. 
Model 1: Edge Acceptance/Rejection

\( w \) accepts the invitation with probability \( p_{vw}(t) \).
Edge Acceptance/Rejection

How can acceptance probability achieve goals of (1) non-power law degree distribution and (2) community structure?

- Rich may choose not to get richer
- Probability of acceptance based on communities
Edge Acceptance/Rejection

How can acceptance probability achieve goals of (1) non-power law degree distribution and (2) community structure?

- Rich may choose not to get richer: $p_{vw}(t) \downarrow 0$
- Probability of acceptance based on communities
Edge Acceptance/Rejection

How can acceptance probability achieve goals of (1) non-power law degree distribution and (2) community structure?

- Rich may choose not to get richer: $p_{vw}(t) \downarrow 0$
- Probability of acceptance based on communities:

$$p_{vw}(t) = \begin{cases} 
    p & C_v = C_w \\
    q & C_v \neq C_w.
\end{cases}$$
Edge Acceptance/Rejection

For now, constant acceptance probability

\[ p_{vw}(t) = p \quad \text{for all } v, w \text{ and } t \geq 0. \]