Large-Scale Networks

3 - Network evolution
Introduction

- We already saw how nodes are linked in networks
  - With what strength they are tied
  - What position makes them more or less important in the network

- Today we will talk about evolution and
  - How the network influences link creation through selection
  - How links influence nodes through social influence
Partitioning
Network partitioning

- Formal definitions are crucial to identify densely connected groups

- *Graph partitioning* consists of breaking a network down into a set of tightly-knit regions with sparser interconnections between the regions
Network partitioning

- Co-authorship of physicists and applied mathematicians working on networks
Network partitioning

- This network contains the tightly-knit groups and weak-ties we observed

- Is there a general way to pull these groups out of the data?
  - **Divisive methods**: some consist of identifying and removing the “spanning links” between densely connected regions, and reiterate on the resulting graph
  - **Agglomerative methods**: others consists of finding nodes that are likely to belong to the same region and merge them together; at the end merged chunks contain the seeds of a densely connected region
Network partitioning

- Consider the following network on the left

- The divisive method identifies nested regions within larger regions by removing edge <7,8> first

- The agglomerative method achieves the same result the other way: by first merging the four triangle into clumps and then pairing off triangle
Network partitioning

- A naive idea is to remove:
  - Local bridges
  - Bridges

- This is insufficient:
  - It does not tell us which bridge to remove first among multiple ones
  - Even if all edges belong to a triangle (no local bridges), a subdivision may be needed
Network partitioning

- However, local bridges form parts of the shortest path between pairs of nodes in different parts of the network
  - Without local bridges paths between some pairs would be re-routed through a longer way
  - We thus look at the edges that carry the most of the traffic. This would be a good candidate for removals

- For each pair of nodes A and B that are connected by a path, we consider 1 unit of flow along the edges from A to B.
  - The flow divides itself evenly along all possible shortest paths
  - If there are k paths from A to B, then 1/k flow passes along each path
Network partitioning

- The *betweenness* of an edge is the total amount of flow it carries
  - We count the flow between all pairs of nodes using this edge

- Consider, for example, edge <7,8>
  - Full flow unit from *pairs of nodes on different sides* of the graph passes through it
  - No flow unit from *pairs of nodes on the same side* passes through it

⇒ The betweenness of this edge is $7 \times 7 = 49$
The Girvan-Newman method

Girvan-Newman method (Input Graph G):

Initially:
i = 1.

Loop:
As long as there are edges in G, repeat:
   Calculate the highest betweenness of G’s edges
   Remove all edges with the highest betweenness from G
   If this splits the graph into new components,
   call them the $i^{th}$ level regions
   increment i

End
The Girvan-Newman method

1. Edge $<5,7>$ carries all traffic from nodes 1-5 to nodes 7-11 for a betweenness of 25. The $<5,6>$ edge only carries flows from 6 to 1-5 for a betweenness of 5 (same for the $<6,7>$ edge)
2. All 25 units of flow that used to be on this deleted edge have shifted onto the path through nodes 5, 6, and 7, and so the betweenness of the 5-6 edge (and also the 6-7 edge) has increased to $5 + 25 = 30$.  

The Girvan-Newman method
The Girvan-Newman method

3. Edges \( <1,2> \), \( <1,3> \), \( <9,11> \), \( <10,11> \) have the highest betweenness hence they are removed next

4. The remaining edges obtain the same betweenness and are removed last
Partitioning of large-scale networks

- This method can be costly to run on a large data set.

- The betweenness quantity requires to reason about all shortest paths between pairs of nodes and should be computed for each remaining edge at each step, which can take very long.

- There is a clever way to compute betweenness based on Breadth-First Search.
Partitioning of large-scale networks

- Consider the graph from one node at a time
- Let see how the flow from this node to all others spread along the edges
- Then, by adding the flows we obtain the betweenness on every edge

Example: consider this graph and how the flow from A reaches all nodes
Partitioning of large-scale networks

1. Perform a breadth-first search of the graph starting at A

2. Determine the number of shortest paths from A to each other node

3. Based on these numbers, determine the amount of flow from A to all other nodes that use each edge
Partitioning of large-scale networks

1. Executing breadth-first search (BFS)
   - Breadth-first search divides a graph into layers with all nodes in layer $d$ at distance $d$ from the source ($A$, in our example).
   - The shortest paths from $A$ to any node is precisely the downward path through increasing layers from $A$ to this node.

   There are two shortest paths from $A$ to $F$, each of length 2:
   - One is $A$, $B$, $F$, the other is $A$, $C$, $F$. 
Partitioning of large-scale networks

2. Determining the number of shortest paths

   - Note that all shortest paths from A to I must take their last step through F or G since these are the two nodes above I in the breadth-first search

   - To be shortest path to I, it must be a shortest path to F or G followed by the corresponding one of these two last edges
2. Determining the number of shortest paths (con’t)

We can generalize this:

- Each node in the first layer is a neighbor of A so it has only one shortest path to A
- We move down through the BFS layers and compute the number of shortest paths as the sum of the number of shortest paths to all nodes directly above it in the BFS.
Partitioning of large-scale networks

3. Determining flow values

1. Let’s start at the bottom layer (with node K):
   A single unit of flow arrives to K and an equal number of shortest paths from A to K go through I and J so this unit of flow is equally divided between \(<I,K>\) and \(<J,K>\)

2. Let’s go upward: the total amount of flow arriving at I is equal to the one unit actually destined for I + the half unit passing through to K = 3/2. How does this get divided over the edges leading upward from I to F and G, respectively? We see that there are twice as many shortest paths from A through F as through G, so twice as much flow should come from F, thus we put 1 unit of flow on F and a half-unit of the flow on G.

3. We continue like this for each other node working upward through the layers
Partitioning of large-scale networks

- To complete the process, we build this BFS structure for each node in the network, determine flow values using this procedure and then sum up the flow values to get the betweenness value for each edge.
- Note that we count the flow values between each pair of nodes X, Y twice: once we do the BFS from X and when we do it from Y. So at the end, we divide everything by 2 to cancel out this double counting.

**Conclusion:** we can identify the edges with the highest betweenness to partition (reasonably) large-scale networks using the Girvan-Newman method.
Homophily
Homophily

› So far, we have discussed the network as a standalone object
› Let’s focus now on the impact of the context in which this network evolves

Homophily is one of the most basic notions impacting social networks
› Your friends do not look like a random sample of the population
› They are generally similar in terms of:
  - Ethnic dimensions
  - Age
  - Mutable characteristics (place they live, occupations, interests, beliefs, opinions)
  - Even though most of us have specific friendships crossing these boundaries
Homophily

› Homophily can divide a social network into densely connected homogeneous parts that are weakly connected to each other.

› In this social network from a town’s middle school and high school divided by ethnicities (left to right) and by schools (top to bottom)

› Can you conclude anything?
Homophily

› How to measure homophily?
  - Is homophily genuinely present in the network?
  - Or is it an artifact of the way the network is drawn?
Homophily

- Consider a friendship network of an elementary-school classroom
  - Assume it exhibits homophily by gender (boys are friends, girls are friends)
  - Consider this graph where pink nodes are girls, white nodes are boys:
  
  ![Graph with pink nodes representing girls and white nodes representing boys](image)

  - If there were no cross-gender edges, it would be easy to identify homophily

- Let's try to define a precise metric of homophily
Homophily

- Assume a node is a boy with probability $p$ and a girl with probability $q$

- Consider a given edge of this network
  - Both ends would be boys with probability $p^2$
  - Both ends would be girls with probability $q^2$
  - A cross-gender edge (with one end a boy and one end a girl) has probability $2pq$

- We can summarize the test of homophily of gender as:

**Homophily test:** if the fraction of cross-gender edges is significantly less than $2pq$, then there is evidence for homophily
Homophily

- Going back to the elementary-school graph

- 5 of the 18 edges in the graph are cross-gender edges
- \( p = \frac{6}{9} = \frac{2}{3} \) and \( q = \frac{3}{9} = \frac{1}{3} \)
- \( 2pq = \frac{4}{9} = \frac{8}{18} \) (one should expect 8 cross-gender edges rather than 5)

⇒ This example seems to show evidence of homophily
Homophily

Remarks
- The number of cross-gender edges in a random assignment of genders will deviate somewhat from its expected value of $2pq$.
- To perform the test in practice, one should quantify the significance of a deviation below a mean.
- A network may have a fraction of cross-gender edges that is significantly more than $2pq$, hence exhibiting inverse homophily (e.g., more opposite sex partners than same-sex partners).
- We can generalize to characteristics taking more than two possible values.
  - An heterogeneous edge connects two different nodes according to this characteristic.
  - We compare the number of heterogeneous edges to their expected number.
Homophily

Mechanisms underlying homophily

- **Selection**: the tendency of people to form friendship with people with similar characteristics
  - Individual characteristics drive the formation of links

- **Social influence**: the tendency of people to change their behavior to be more closely into alignment with the behavior of their friends
  - The existing links shape people’s characteristics

⇒ **Social influence** can be considered as the reverse of selection
Affiliation
Affiliations

- **Foci** are focal points of social interactions constituting social, psychological, legal or physical entities around which joint activities are organized (e.g., workplaces, voluntary organizations, hangouts, etc.) [Feld81]

- We can represent the participation of people in a set of foci as an affiliation network or a *bipartite graph* whose nodes can be divided into two sets such that every edge connects a node in one set to a node in the other set.
Affiliations

- Affiliation networks were used to study the composition of boards of directors of major corporations [Miz96]
  - Many board members serve on multiple boards

NB. This affiliation network is from mid-2009.
Affiliations

› **Social networks and affiliation networks** coevolve with one another
  - Participation in a shared focus gives opportunity for friendship

› Let us define a **social-affiliation network** with two kinds of edges:
  - A social one linking two people if they are friends
  - An affiliation one linking a person to a focus

A social-affiliation network shows both the friendships between people and their affiliation with different social foci
**Link formation**

- Different edge creations in social-affiliation network are **closure** processes

1. **Triadic closure**: all nodes are persons

2. **Focal closure**: two people being linked under a common focus influence

3. **Membership closure**: a new affiliation being created under the friendship influence
Link formation

- **Examples:**
  1. Bob introduces Anna to Claire (triadic closure)
  2. Karate introduces Anna to Daniel (focal closure)
  3. Anna introduces Bob to Karate (membership closure)
Link formation

How to evaluate empirically triadic closure in a dynamic network?

- Get an estimate of the probability of friendship formation $T(k)$ under the effect of $k$ common friends

Empirical estimation of the probability of link formation under the effect of common friends

1. Take 2 snapshots of the network at different times

2. For each $k$, identify all pairs of nodes who have exactly $k$ friends in common in the first snapshot, but who are not directly connected by an edge.

3. $T(k) = \text{the fraction of these pairs that have formed an edge by the time of the second snapshot}$

- Plot $T(k)$ as a function of $k$ to illustrate the effect of common friends on the formation of links
Link formation

Anna and Esther have two friends in common, while Claire and Daniel only have one friend in common.

How much more likely is the formation of a link in the first of these two cases?
Kossinets and Watts computed $T(k)$ on a dataset of email communication among 22,000 undergraduate and graduate students over a one-year period at a large US university [KW06]. They linked two people at some instant if they had exchanged an email in the last 60 days and computed $T(k)$ on each pair of one-day apart snapshots. They average all the curves to obtain the average probability that two people form a link per day as a function of the number of common friends they have.
Link formation

Kossinets and Watts computed $T(k)$ on a dataset of email communication among 22000 undergraduate and graduate students over a one-year period at a large US university [KW06]. They linked two people at some instant if they had exchanged an email in the last 60 days and computed $T(k)$ on each pair of one-day apart snapshots. They average all the curves to obtain the average probability that two people form a link per day as a function of the number of common friends they have.

$T(0)$ close to 0

They linked two people at some instant if they had exchanged an email in the last 60 days and computed $T(k)$ on each pair of one-day apart snapshots. They average all the curves to obtain the average probability that two people form a link per day as a function of the number of common friends they have.
Link formation

Kossinets and Watts computed $T(k)$ on a dataset of email communication among 22000 undergraduate and graduate students over a one-year period at a large US university [KW06].

They linked two people at some instant if they had exchanged an email in the last 60 days and computed $T(k)$ on each pair of one-day apart snapshots. They average all the curves to obtain the average probability that two people form a link per day as a function of the number of common friends they have.
Link formation

Kossinets and Watts computed $T(k)$ on a dataset of email communication among 22000 undergraduate and graduate students over a one-year period at a large US university [KW06].

They linked two people at some instant if they had exchanged an email in the last 60 days and computed $T(k)$ on each pair of one-day apart snapshots. They average all the curves to obtain the average probability that two people form a link per day as a function of the number of common friends they have.

Significant increase at 8-10 but on a relatively small sample of the population: Few people have 8-10 friends in common without already having formed a link.
Link formation

Simple baseline model

- Suppose that for some small probability $p$,
  - Each common friend that two people have give them an independent probability $p$ of forming a link each day
  - So if two people have $k$ friends in common, the probability that they fail forming a link is $(1-p)^k$
  - The probability that a link forms each day for a given pair of persons with $k$ friends in common is $1-(1-p)^k$ (this corresponds to the left dashed line)
  - Given the small absolute effect of the first common friend in the data, we also do a comparison to the curve $1-(1-p)^{k-1}$, which shifts the baseline curve one unit to the right (this corresponds to the right dashed line)

- A larger and more detailed study of these effects was conducted by Leskovec et al. on LinkedIn, Flickr, Del.icio.us and Yahoo! Answers [LBKT08]
Wikipedia as an Example
Effect of selection and social influence

**Wikipedia** is an updatable online encyclopedia

- Wikipedia consists of a set of *pages*, each providing information on a particular topic
- The text in a page may link to other pages on related topics through **hyperlinks**
- **Editors** are people who edit Wikipedia pages
- Editors have a user account and a *user talk page* where someone else can leave a message for the editor
- Every action (edit, message) on Wikipedia is recorded and timestamped
Effect of selection and social influence

- Consider the Wikipedia social-affiliation network [CCH+08]
  - Each node is a Wikipedia editor with a maintained user account and user talk page on the system
  - An edge joins two editors if one has written on the user talk page of the other
  - The foci are Wikipedia articles
  - There is an affiliation between an editor and a focus if it has edited the corresponding article
Effect of selection and social influence

The probability that a person edits a Wikipedia article as a function of the number of prior editors of that article with whom he or she has communicated.

The probability increases with the number \( k \) of common neighbors, representing friends associated with the foci.
Effect of selection and social influence

How selection and social influence work together to produce homophily?

- Let’s define behavior similarity
  - The editor’s behavior corresponds to the set of article she has edited
  - So let’s consider the bipartite affiliation network:
    - Where nodes are editors and articles
    - And links are from editors to articles they have edited
  - A simple behavior similarity definition is thus the neighborhood overlap (seen last week) of this resulting bipartite affiliation network
  - More generally we can define behavior similarity as:

\[
\text{number of articles edited by both } A \text{ and } B \\
\text{number of articles edited by at least one of } A \text{ or } B
\]
Effect of selection and social influence

Quantifying the interplay between selection and social influence

- For each pair of editors A and B who have ever communicated, record their similarity over time, where “time” in this case moves in discrete time units, advancing by one “tick” whenever either A or B edits or sends messages.

- Next, declare time 0 for the pair <A,B> when A and B first communicated.

- We obtain many curves showing similarities as a function of time.

- All these curves can be averaged into the level of similarity as a function of time of first interaction, over all pairs of editors who have ever interacted.
Effect of selection and social influence

- Similarity increases before and after the first interaction
  \[\Rightarrow\] Both selection and social influence are at work

- Non-symmetric curve: selection raises similarity rapidly
Spatial Model
Spatial model of segregation

Percentage of African Americans per city block of Chicago for the year 1940-1960 (lower is lighter): concentration intensifies with time.
Spatial model of segregation

The Schelling Model [Sch78]
- We assume a population of individuals of either immutable type O or X, these are *agents*.
- A grid is a 2-dimensional representation of a city.
- Each agent resides in a cell of this grid, each having 8 neighbor cells.
- Below a threshold $t$ of similar neighbors, an agent moves to another cell.

Example: right fig. shows with * unsatisfied people from the left fig. w/ $t=3$.

<table>
<thead>
<tr>
<th>X</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>O</td>
</tr>
<tr>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X1*</th>
<th>X2*</th>
</tr>
</thead>
<tbody>
<tr>
<td>X3</td>
<td>O1*</td>
</tr>
<tr>
<td>O2</td>
<td>O3</td>
</tr>
<tr>
<td>X4</td>
<td>X5</td>
</tr>
<tr>
<td>O4</td>
<td>O5*</td>
</tr>
<tr>
<td>X6*</td>
<td>O6</td>
</tr>
<tr>
<td>X7</td>
<td>X8</td>
</tr>
<tr>
<td>O7</td>
<td>O8</td>
</tr>
<tr>
<td>X9*</td>
<td>X10</td>
</tr>
<tr>
<td>X11</td>
<td>O11*</td>
</tr>
</tbody>
</table>
Spatial model of segregation

Agent movements
- Agents move in a sequence of rounds
- In each round we move the unsatisfied agents in turn to a satisfying cell

Example: agents from left figure are moved, one at time from top-left to bottom-right, to the nearest satisfying cell, resulting in the right figure
Spatial model of segregation

- Model simulated on 10000 agents
  - Grid of 150x150 cells
  - Threshold $t = 3$
  - Unsatisfied agents move to a random cell

- Example: satisfaction states reached after $\sim 50$ iterations from two different starting states
- The model produces large homogeneous regions, interlocking with each other as they stretch across the grid
Homophily

- With threshold $t=4$, the segregation is more intense

- In the same settings as before, we report the state after:
  - 20 rounds
  - 150 rounds
  - 350 rounds
  - 800 rounds

- We obtain a single significant region of each type
References


