Graph Theory and Social Networks - part 5

EE599: Social Network Systems

Keith M. Chugg
Fall 2014
Overview

- Last part of deterministic graph models for social networks
  - Finish partitioning
  - Segregation, polarization (Easley & Kleinberg, Ch. 4, 5)
    - Context for social networks
    - Homophily
    - Positive/negative links
Network Partitioning Approaches

- Multiple approaches [Jackson 13.2]
  - Edge removal - e.g., Girvan-Neman
  - Hierarchical Clustering
    - Group like nodes according to some similarity measure
  - Iterated correlation (CONCOR)
    - Iterate a correlation measure on rows of adjacency matrix
  - Maximum Likelihood approaches (EM algorithm)
    - Find partition that maximizes probability of observed graph given intra-community and inter-community edge probabilities
Network Partitioning

- When to stop?
  - Need a measure of how good a partition is...
  - One measure is \textit{modularity} — compare how much larger intra-community edge frequencies with those that would be expected under random connections

- Example: Girvan-Newman
  - \textit{stop removing edges when modularity decreases}
  - \textit{go through entire process and select partition with highest modularity}
Modularity

- Measures how much larger the fraction of intra-community edges is relative to the fraction expected under random placement
  - Positive modularity => captures community structure
  - Zero modularity => same as expected under random grouping
  - Negative modularity => worse than random grouping
Modularity - Random Model

Randomly select from the $2L$ edge stubs to connect with the $L$ edges.

$$
\text{Prob}(\text{selecting an edge stub connected to node } i) = \frac{k_i}{2L}
$$

$$
\text{Prob}(\text{selecting an edge stub connected to node } j) = \frac{k_j}{2L}
$$

$$
\text{Prob}(\text{randomly connecting } i \text{ & } j) = \frac{k_i k_j}{4L^2}
$$

(This may have some issues since it allows for nodes to connect to themselves)
Modularity

\[ Q = \sum_{m=1}^{M} \frac{a_{ij}}{2L} \sum_{i,j \in C_m} - \sum_{m=1}^{M} \frac{k_i k_j}{2L} \sum_{i,j \in C_m} \frac{1}{2L} \sum_{i \in C_m} k_i = \sum_{m=1}^{M} e_{mm} - r_m^2 \]

where
- \( Q \) is the modularity
- \( a_{ij} \) is the fraction of all edges connecting nodes in the same community
- \( k_i k_j \) is the fraction of all edges connecting nodes in community \( m \) to other nodes in community \( m \)
- \( e_{mm} \) is the fraction of all edges connecting nodes in the same community under a random model
- \( r_m \) is the fraction of edges connecting nodes in the same community
- \( \sum_{i \in C_m} k_i \) is the sum over all communities
I suspect that there may be some other reasonable (better?) ways to define the quality of a partition.
Network Partitioning & Betweenness

- Community Detection/Partitioning

- Brandes Algorithm [77] in Easley & Kleinberg

- Code available to compute many of these properties and algorithms
  - MatLabBGL (www.mathworks.com/matlabcentral/fileexchange/10922-matlabbgl)

- Modularity & Girvan-Newman
Overview

- Last part of deterministic graph models for social networks
  - Finish partitioning
  - Segregation, polarization \((Easley & Kleinberg, \text{Ch. 4, 5})\)
    - Context for social networks
    - Homophily
    - Positive/negative links
Homophily

- *Homophily* is the property of “like” nodes to connect in a social network
- Like ~ race, interest, age, etc.
- Two predominant mechanisms
  - *Selection*: people tend to befriend people similar to themselves
  - *Social Influence* (peer pressure): once in with a group, people tend to conform
Figure 4.1: Homophily can produce a division of 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, two such divisions in the network are apparent: one based on race (with students of different races drawn as differently colored circles), and the other based on friendships in the middle and high schools respectively [304].

Easley & Kleinberg
Homophily Test

Figure 4.2: Using a numerical measure, one can determine whether small networks such as this one (with nodes divided into two types) exhibit homophily.

- Expected inter-type link frequency under random connections: \(2pq = 4/9 = 8/18\)
- Observed inter-type links frequency: 5/18

“some evidence of homophily”
Homophily Test

- White nodes occur with freq: $p = \frac{2}{3}$
- Pink nodes occur with freq: $q = 1 - p = \frac{1}{3}$

- Expected frequencies for random connections:
  - white-pink/pink-white $2pq$

Homophily Measure $\sim \frac{5}{18} - 2\left(\frac{2}{3}\right)\left(\frac{1}{3}\right) = -0.17 < 0$
Social Network Context

Figure 4.4: One type of affiliation network that has been widely studied is the memberships of people on corporate boards of directors [301]. A very small portion of this network (as of mid-2009) is shown here. The structural pattern of memberships can reveal subtleties in the interactions among both the board members and the companies.

- **Affiliation Network**: what foci are people members of?

Easley & Kleinberg
Social Network Context

- Social network and can be combined with affiliations to provide context

Figure 4.5: A social-affiliation network shows both the friendships between people and their affiliation with different social foci.  

Easley & Kleinberg
4.4 Tracking Link Formation in On-Line Data

In this chapter and the previous one, we have identified a set of different mechanisms that lead to the formation of links in social networks. These mechanisms are good examples of social phenomena which are clearly at work in small-group settings, but which have traditionally been very hard to measure quantitatively. A natural research strategy is to try tracking these mechanisms as they operate in large populations, where an accumulation of many small effects can produce something observable in the aggregate. However, given that most of the forces responsible for link formation go largely unrecorded in everyday life, it is a challenge to select a large, clearly delineated group of people (and social foci), and accurately quantify the relative contributions that these different mechanisms make to the formation of real network links.

The availability of data from large on-line settings with clear social structure has made it possible to attempt some preliminary research along these lines. As we emphasized in Chapter 2, any analysis of social processes based on such on-line datasets must come with a number of caveats. In particular, it is never a priori clear how much one can extrapolate from digital interactions to interactions that are not computer-mediated, or even from one computer-mediated setting to another. Of course, this problem of extrapolation is present whenever one studies phenomena in a model system, on-line or not, and the kinds of measurements these large datasets enable represent interesting first steps toward a deeper quantitative understanding of how mechanisms of link formation operate in real life. Exploring these questions in a broader range of large datasets is an important problem, and one that will become easier as large-scale data becomes increasingly abundant.

Triadic closure.

With this background in mind, let’s start with some questions about triadic closure. Here’s a first, basic numerical question: how much more likely is a link to
Figure 4.9: Quantifying the effects of triadic closure in an e-mail dataset [259]. The curve determined from the data is shown in the solid black line; the dotted curves show a comparison to probabilities computed according to two simple baseline models in which common friends provide independent probabilities of link formation.

Easley & Kleinberg

- Triadic closure as a function of number of common friends

\[ 1 - (1 - p)^k \]
\[ 1 - (1 - p)^{k-1} \]
Examples from Data

Figure 4.10: Quantifying the effects of focal closure in an e-mail dataset [259]. Again, the curve determined from the data is shown in the solid black line, while the dotted curve provides a comparison to a simple baseline.

- Foci closure as a function of number of common foci
- Lesser effect of many shared foci as compared to many shared friends
Examples from Data

- Membership closure as a function of number of common memberships in online communities
- Lesser effect of many shared foci as compared to many shared friends

Figure 4.11: Quantifying the effects of membership closure in a large online dataset: The plot shows the probability of joining a LiveJournal community as a function of the number of friends who are already members [32].

Easley & Kleinberg
Examples from Data

- Membership closure as a function of number of common pages edited on Wikipedia
- Lesser effect of many shared foci as compared to many shared friends

Figure 4.12: Quantifying the effects of membership closure in a large online dataset: The plot shows the probability of editing a Wikipedia article as a function of the number of friends who have already done so [122].

Easley & Kleinberg
Examples from Data

- Many shared friends is the strongest force for closure

- In all cases, there is a significant jump in closure rates from having 1 entity in common to having 2 entities in common
Relative Cause of Homophily

\[
\frac{\text{number of articles edited by both } A \text{ and } B}{\text{number of articles edited by at least one of } A \text{ or } B}
\]

- Selection is a stronger cause for homophily in this experiment

Figure 4.13: The average similarity of two editors on Wikipedia, relative to the time (0) at which they first communicated \[122\]. Time, on the x-axis, is measured in discrete units, where each unit corresponds to a single Wikipedia action taken by either of the two editors. The curve increases both before and after the first contact at time 0, indicating that both selection and social influence play a role; the increase in similarity is steepest just before time 0.

Easley & Kleinberg
Shelling’s Model for Segregation

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

(a) Agents occupying cells on a grid.

(b) Neighbor relations as a graph.

Figure 4.15: In Schelling’s segregation model, agents of two different types (X and O) occupy cells on a grid. The neighbor relationships among the cells can be represented very simply as a graph. Agents care about whether they have at least some neighbors of the same type.

- People require that \( t \) of their 8 neighbors are like them or else they move
Shelling's Model for Segregation

If a node is not satisfied (<t like neighbors), then she moves to an empty spot

Figure 4.16: After arranging agents in cells of the grid, we first determine which agents are unsatisfied, with fewer than t other agents of the same type as neighbors. In one round, each of these agents moves to a cell where they will be satisfied; this may cause other agents to become unsatisfied, in which case a new round of movement begins.

Easley & Kleinberg
Shelling’s Model for Segregation

- t=3 condition can be satisfied, but these are not likely to be found by randomly moving when dissatisfied.

---

Figure 4.18: With a threshold of 3, it is possible to arrange agents in an integrated pattern: all agents are satisfied, and everyone who is not on the boundary on the grid has an equal number of neighbors of each type.

Easley & Kleinberg
Shelling’s Model for Segregation

- t=3 is “tolerant” at the level of the individual, but still results in segregation with very high probability!

Figure 4.17: Two runs of a simulation of the Schelling model with a threshold t of 3, on a 150-by-150 grid with 10,000 agents of each type. Each cell of the grid is colored red if it is occupied by an agent of the first type, blue if it is occupied by an agent of the second type, and black if it is empty (not occupied by any agent).

Easley & Kleinberg
Shelling’s Model for Segregation

Figure 4.19: Four intermediate points in a simulation of the Schelling model with a threshold \( t \) of 4, on a 150-by-150 grid with 10,000 agents of each type. As the rounds of movement progress, large homogeneous regions on the grid grow at the expense of smaller, narrower regions.

- \( t=4 \) is even more pronounced
Shelling’s Model for Segregation

- Many on-line simulators for this and related models
  - e.g., http://nifty.stanford.edu/2014/mccown-schelling-model-segregation/
Overview

- Last part of deterministic graph models for social networks
  - Finish partitioning
  - Segregation, polarization (Easley & Kleinberg, Ch. 4, 5)
    - Context for social networks
    - Homophily
  - Positive/negative links
Positive & Negative Relationships

• Not all connections are positive friendships

• What can we learn from modeling interactions as either positive (friends) or negative (enemies)?
Positive & Negative Relationships

(a) A, B, and C are mutual friends: balanced.

(b) A is friends with B and C, but they don’t get along with each other: not balanced.

(c) A and B are friends with C as a mutual enemy: balanced.

(d) A, B, and C are mutual enemies: not balanced.

Figure 5.1: Structural balance: Each labeled triangle must have 1 or 3 positive edges.

Easley & Kleinberg
Structural Balance

Figure 5.2: The labeled four-node complete graph on the left is balanced; the one on the right is not.

Easley & Kleinberg

• Balanced network = all triangles have 3 or 1 +’s
Balanced Complete Networks

Claim: If all triangles in a labeled complete graph are balanced, then either
(a) all pairs of nodes are friends, or else
(b) the nodes can be divided into two groups, X and Y, such that
   (i) every pair of nodes in X like each other,
   (ii) every pair of nodes in Y like each other, and
   (iii) everyone in X is the enemy of everyone in Y.

Figure 5.3: If a complete graph can be divided into two sets of mutual friends, with complete mutual antagonism between the two sets, then it is balanced. Furthermore, this is the only way for a complete graph to be balanced.

Easley & Kleinberg

- Balance Theorem: any **complete** balanced network has the above structure — pure two-party polarization. Other option is that everybody is friends (Y is empty set)
Proof of Balance Theorem

Figure 5.4: A schematic illustration of our analysis of balanced networks. (There may be other nodes not illustrated here.)

Easley & Kleinberg

node A selected arbitrarily
Proof of Balance Theorem

Figure 5.4: A schematic illustration of our analysis of balanced networks. (There may be other nodes not illustrated here.)

Easley & Kleinberg

- Other option is all
Example: Alliance Evolution Leading to WWI

Figure 5.5: The evolution of alliances in Europe, 1872-1907 (the nations GB, Fr, Ru, It, Ge, and AH are Great Britain, France, Russia, Italy, Germany, and Austria-Hungary respectively). Solid dark edges indicate friendship while dotted red edges indicate enmity. Note how the network slides into a balanced labeling — and into World War I. This figure and example are from Antal, Krapivsky, and Redner [20].

Easley & Kleinberg
Generalizations of Structural Balance

- Approximately balanced (complete) networks
- Structural balance for incomplete graphs
  - Definition and algorithmic test
- Weak balance (complete networks)
  - Only prohibit 2 + per triangle
  - Yields multiparty, mutually antagonistic structure
Approximately Balanced Complete Networks

general statement

Claim: Let $\varepsilon$ be any number such that $0 \leq \varepsilon < \frac{1}{5}$, and define $\delta = \sqrt{\varepsilon}$. If at least $1 - \varepsilon$ of all triangles in a labeled complete graph are balanced, then either

(a) there is a set consisting of at least $1 - \delta$ of the nodes in which at least $1 - \delta$ of all pairs are friends, or else

(b) the nodes can be divided into two groups, $X$ and $Y$, such that

(i) at least $1 - \delta$ of the pairs in $X$ like each other,

(ii) at least $1 - \delta$ of the pairs in $Y$ like each other, and

(iii) at least $1 - \delta$ of the pairs with one end in $X$ and the other end in $Y$ are enemies.

Easley & Kleinberg

• Almost all triangles are balanced implies almost completely segregated

Claim: If at least 99.9% of all triangles in a labeled complete graph are balanced, then either

(a) there is a set consisting of at least 90% of the nodes in which at least 90% of all pairs are friends, or else

(b) the nodes can be divided into two groups, $X$ and $Y$, such that

(i) at least 90% of the pairs in $X$ like each other,

(ii) at least 90% of the pairs in $Y$ like each other, and

(iii) at least 90% of the pairs with one end in $X$ and the other end in $Y$ are enemies.
Approximately Balanced Complete Networks

- Outline of proof

- Find a “good” node A

- Find average number of unbalanced triangles in which nodes are involved (combinatorics)

- Pigeon-hole principle implies that there is at least one node involved in at most this average number

- Split network into friends/enemies of this good node A

- Bound number of unbalanced conditions inside friend group, enemy group, and between these groups

Figure 5.17: The characterization of approximately balanced complete graphs follows from an analysis similar to the proof of the original Balance Theorem. However, we have to be more careful in dividing the graph by first finding a “good” node that isn’t involved in too many unbalanced triangles.

Easley & Kleinberg
Generalizations of Structural Balance

- Approximately balanced (complete) networks
- Structural balance for incomplete graphs
  - Definition and algorithmic test
- Weak balance (complete networks)
  - Only prohibit 2+ per triangle
  - Yields multiparty, mutually antagonistic structure
Structural Balance in Incomplete Networks

Is this network balanced?

Yes: can fill in missing edges consistent with a complete, balanced network

Yes: can partition nodes in two mutually agnostic groups of friends

Figure 5.9: There are two equivalent ways to define structural balance for general (non-complete) graphs. One definition asks whether it is possible to fill in the remaining edges so as to produce a signed complete graph that is balanced. The other definition asks whether it is possible to divide the nodes into two sets \( X \) and \( Y \) so that all edges inside \( X \) and inside \( Y \) are positive, and all edges between \( X \) and \( Y \) are negative.

Easley & Kleinberg

- These two “definitions” are equivalent — this is structural balance in incomplete networks
Structural Balance in Incomplete Networks

Is this network balanced?

Figure 5.8: In graphs that are not complete, we can still define notions of structural balance when the edges that are present have positive or negative signs indicating friend or enemy relations.

- Difficult to ascertain as networks becomes nontrivial
Structural Balance in Incomplete Networks

- Difficult to ascertain as networks becomes nontrivial

Figure 5.8: In graphs that are not complete, we can still define notions of structural balance when the edges that are present have positive or negative signs indicating friend or enemy relations.

Easley & Kleinberg
Can we just start labeling a node in one group and change group label when we encounter a negative connection?
Can we just start labeling a node in one group and change group label when we encounter a negative connection?
• Can we just start labeling a node in one group and change group label when we encounter a negative connection?
Can we just start labeling a node in one group and change group label when we encounter a negative connection?
• Can we just start labeling a node in one group and change group label when we encounter a negative connection?
Can we just start labeling a node in one group and change group label when we encounter a negative connection?
Can we just start labeling a node in one group and change group label when we encounter a negative connection?
Can we just start labeling a node in one group and change group label when we encounter a negative connection?
Figure 5.10: If a signed graph contains a cycle with an odd number of negative edges, then it is not balanced. Indeed, if we pick one of the nodes and try to place it in $X$, then following the set of friend/enemy relations around the cycle will produce a conflict by the time we get to the starting node.

Easley & Kleinberg
Structural Balance in Incomplete Networks

- A signed graph (possibly incomplete) is structurally balanced if and only if it contains no cycles with an odd number of negative edges

- Proof outline:
  - Create a reduced graph with components containing only positive nodes reduced to a “super node”
    - This reduced graph has only negative edges
  - Original graph is balanced iff reduced graph is bipartite
  - Use BFS to test if the reduced graph is bipartite
Proof: Balance in Incomplete Networks

• Try to obtain a “balanced division” — i.e., a partition of nodes into groups X & Y such that all are friends inside X and inside Y, but all edges between X and Y are negative

• First, find the connected components if only positive edges are considered

• These are candidates for groups of X and Y labels (i.e., groups of friends) — “positive blobs”

• If there is a negative edge in any of these positive blobs, the network is not balanced (see next slide)
Proof: Balance in Incomplete Networks

Figure 5.11: To determine if a signed graph is balanced, the first step is to consider only the positive edges, find the connected components using just these edges, and declare each of these components to be a **supernode**. In any balanced division of the graph into \( X \) and \( Y \), all nodes in the same supernode will have to go into the same set.

---

- Creating positive blobs by considering connected components when negative edges are removed

---

© Keith M. Chugg, 2014
• Positive blob cannot have a negative edge. A negative edge makes it impossible to label all nodes in the blob the same. It creates a cycle with an odd number of negative edges.

• If negative edge in positive blob, then network is unbalanced
Proof: Balance in Incomplete Networks

- If no negative edges in positive blobs, create a reduced graph with:
  - Positive blobs become “super nodes” (all mutual friends, so all labeled the same)
  - All edges are negative
Proof: Balance in Incomplete Networks

- Problem now is whether super nodes can be labeled X and Y so that neighbors differ. This is possible:
  - iff no odd length cycles in reduced graph
  - iff no cycles in original graph with odd number of negative edges
  - iff reduced graph has a bipartite labeling

Figure 5.14: A more standard drawing of the reduced graph from the previous figure. A negative cycle is visually apparent in this drawing.

The second possible outcome will be to find a cycle in the reduced graph that has an odd number of edges. We can then convert this to a (potentially longer) cycle in the original graph with an odd number of negative edges: the cycle in the reduced graph connects supernodes, and corresponds to a set of negative edges in the original graph. We can simply "stitch together" these negative edges using paths consisting entirely of positive edges that go through the insides of the supernodes. This will be a path containing an odd number of negative edges in the original graph.

For example, the odd-length cycle in Figure 5.14 through nodes A through E can be realized in the original graph as the darkened negative edges shown in Figure 5.15. This can then be turned into a cycle in the original graph by including paths through the supernodes – in this example using the additional nodes 3 and 12.

In fact, this version of the problem when there are only negative edges is known in graph theory as the problem of determining whether a graph is bipartite: whether its nodes can be divided into two groups (in this case X and Y) so that each edge goes from one group to the other. We saw bipartite graphs when we considered a affiliation networks in Chapter 4, but there the fact that the graphs were bipartite was apparent from the ready-made division of the nodes into people and social foci. Here, on the other hand, we are handed a graph "in reduced graph"
Proof: Balance in Incomplete Networks

- Test for bipartite labeling via BFS in the reduced graph

Figure 5.16: When we perform a breadth-first search of the reduced graph, there is either an edge connecting two nodes in the same layer or there isn’t. If there isn’t, then we can produce the desired division into \(X\) and \(Y\) by putting alternate layers in different sets. If there is such an edge (such as the edge joining \(A\) and \(B\) in the figure), then we can take two paths of the same length leading to the two ends of the edge, which together with the edge itself forms an odd cycle.

Easley & Kleinberg
Generalizations of Structural Balance

- Approximately balanced (complete) networks
- Structural balance for incomplete graphs
  - Definition and algorithmic test
- Weak balance (complete networks)
  - Only prohibit 2 + per triangle
  - Yields multiparty, mutually antagonistic structure
Weakly Balanced Networks

(a) A, B, and C are mutual friends: balanced.

(b) A is friends with B and C, but they don’t get along with each other: not balanced.

(c) A and B are friends with C as a mutual enemy: balanced.

(d) A, B, and C are mutual enemies: not balanced.

Figure 5.1: Structural balance: Each labeled triangle must have 1 or 3 positive edges.

Easley & Kleinberg

Weakly balanced: only disallow triangles with 2 +’s and 1 -
Weakly Balanced Complete Networks

Balance Theorem: any complete weakly balanced network has the above structure — pure multi-party polarization. Other option is that everybody is friends.

Figure 5.6: A complete graph is weakly balanced precisely when it can be divided into multiple sets of mutual friends, with complete mutual antagonism between each pair of sets. Easley & Kleinberg
Proof of Balance Theorem

Figure 5.7: A schematic illustration of our analysis of weakly balanced networks. (There may be other nodes not illustrated here.)

Easley & Kleinberg
Proof of Balance Theorem

Figure 5.7: A schematic illustration of our analysis of weakly balanced networks. (There may be other nodes not illustrated here.)

Easley & Kleinberg

- Apply the same reasoning to the enemies of A to get the result