Homophily and Affiliation
Web Science (VU) (707.000)

Elisabeth Lex
KTI, TU Graz

April 27, 2015
Outline

1. Repetition
2. Homophily
3. Homophily: Selection and Social Influence
4. Affiliation and Affiliation Networks
5. Link Formation in Social Affiliation Networks
6. The Schelling Model
7. Summary and Practical Examples
We discussed structural characteristics of social networks, and some processes that affect formation of links in a network i.e. Triadic Closure

- When individuals B and C have a common friend A, then there are increased opportunities and sources of trust on which to base their interactions, and A will also have incentives to facilitate their friendship.
Motivation (1/2)

- However: without regard of context in which networks form!
- These contexts have effects on network
- Each person in a social network has a distinctive set of characteristics
- Similarities or compatibilities among two people’s characteristics can strongly influence whether a link forms between them
Network’s surrounding contexts: factors that exist outside the nodes and edges of a network and which affect how network evolves

Today, we will discuss how such effects operate, and what they imply about the structure of social networks

We will find that these surrounding contexts can be viewed in network terms as well by expanding the network to represent the contexts together with the individuals

We will discuss different processes of network formation

And a model for segregation in networks based on certain factors
Homophily
Homophily is the principle that we tend to be similar to our friends
E.g. to places to live, occupations, interests, beliefs, and opinions
One of the most basic notions governing the structure of social networks
Most of us have specific friendships that cross such boundaries
But in aggregate, many links in social networks tend to connect people who are similar to one another
“Birds of a feather flock together”
Homophily: Introduction (2/2)

- Compare: friendship forms because two people are introduced through a common friend vs friendship forms because two people attend the same school:
  - First is intrinsic in the network (triadic closure)
  - Second can only be understood when investigating contextual information beyond the network (social environment)
Example: Homophily depicted in a Social Network

- Social network from a town’s middle school and high school
- Two divisions: (i) based on race (students of different races drawn as colored circles, left to right); (ii) based on age and school attendance (students in middle school separated from students in high school, top to bottom)

Homophily can divide social network into densely connected, homogeneous parts that are weakly connected
Homophily and Triadic Closure

- Social contexts provide a basis for triadic closure!
- Why? Say A-B and A-C friendships exist, homophily suggests that B and C are each likely to be similar to A in a number of dimensions, and thus possibly also similar to each other
- Higher chance that B-C friendship will form due to this similarity
- Even if neither of them is aware that the other one knows A
- Ergo: links may arise from a combination of several factors, partly due to the effect of other nodes in the network, and partly due to the surrounding contexts
Measuring Homophily: Example Network with Male and Female People

Suppose we have a network in which a $p$ fraction of all individuals are male, and a $q$ fraction of all individuals are female.

- Consider: random network $R = (V, E_r)$, each node is assigned male with probability $p$, and female with probability $q$
- Let $G = (V, E)$ be a random sample of $R$ with $p$ fraction of male, and $q$ fraction of female.
- Consider any edge $(i, j) \in E_r$ of this random network $R$
  - Let random variable $X_{ij} = 1$ if it is a cross-edge (i.e., first end of the edge is male and the second end is female, or vice versa), and $X_{ij} = 0$ otherwise
  - Then $X_{ij}$ is a Bernoulli random variable such that $P(X_{ij} = 1) = 2pq$
- If fraction of cross-edges is significantly less than $2pq$, then there is evidence for homophily
**Example**

- Shaded nodes are girls, unshaded nodes are boys
- 5 of the 18 edges are cross-gender, so \( p = \frac{2}{3} \) and \( q = \frac{1}{3} \)
- Compare fraction of cross-gender edges to \( 2pq = \frac{4}{9} = \frac{8}{18} \)
- With no homophily, we expect 8 cross-gender edges and not 5 - evidence of homophily.
- In practice, working definition of “significantly less than.”, e.g. statistical significance
- Also possible for network to have fraction of cross-gender edges that is significantly more than \( 2pq \), then network exhibits inverse homophily
Homophily: Selection and Social Influence
Mechanisms Underlying Homophily: Selection and Social Influence (1/2)

- **Selection**: the tendency of people to form friendships with others who are like them in terms of immutable characteristics such as ethnicity.
- Selection may operate at several different levels, and with different levels of intentionality.
- E.g. People choose friends who are most similar from a small group of available persons.
- Globally, selection can be more implicit: For example, when people live in neighborhoods, attend schools, or work for companies that are relatively homogeneous compared to the population at large, the social environment is already favoring opportunities to form friendships with others like oneself.
- With selection, the individual characteristics drive the formation of links.
People may modify their behaviors to bring them more closely into alignment with the behaviors of their friends.

Process has been variously described as socialization and social influence since existing social connections in a network are influencing individual characteristics of nodes.

With social influence, the existing links in the network serve to shape people’s (mutable) characteristics.
Interplay between Selection and Social Influence

- Requires longitudinal studies, in which social connections and behaviors within a group are both tracked over a period of time.
- Enables to see behavioral changes that occur after changes in an individual’s network connections, as opposed to changes to the network that occur after an individual changes his or her behavior.
- E.g. study processes why friends have similar outcomes in terms of scholastic achievement and delinquent behavior such as drug use.
- Hard to resolve is how these two effects interact, and whether one is more strongly at work than the other.
Christakis and Fowler studied effect of social networks on health-related outcomes.

Using longitudinal data covering roughly 12,000 people, they tracked obesity status and social network structure over a 32-year period.

They found that obese and non-obese people clustered in the network in a fashion consistent with homophily: people tend to be more similar in obesity status to their network neighbors than in a version of the same network where obesity status is assigned randomly.
Example of Interplay between Selection and Social Influence (2/2)

Why?

- Because of selection effects, in which people are choosing to form friendships with others of similar obesity status?
- Because of the effects of homophily according to other characteristics, in which the network structure indicates existing patterns of similarity in other dimensions that correlate with obesity status?
- Because changes in the obesity status of a person’s friends was exerting a (presumably behavioral) influence that affected his or her future obesity status?

Christakis and Fowler found that, even accounting for effects of types (i) and (ii), there is significant evidence for an effect of type (iii) as well: that obesity is a health condition displaying a form of social influence, with changes in your friends’ obesity status having a subsequent effect on you.
Summary

- We have discussed contextual factors that affect formation of links in a network
- Based on similarities in characteristics of nodes, behaviours, activities nodes are engaged in
- Careful analysis needed to distinguish between factors that contribute to a conclusion
- May be not clear why people tend to be similar to neighbours and how this affects evolution of network and how behaviour of people in network is influenced
Affiliation and Affiliation Networks
Affiliation

- Any context can be modeled into network itself even if they exist outside through networks that contain both people and contexts as nodes.
- Today: person and activities they participate in and how they affect formation of links.
- Activity: e.g. part of a company, organization, neighbourhood; pursuing a particular hobby or interest.
- Activities when shared between two people tend to increase likelihood that they will interact and form a link in the network.
- “Focal points of social interactions” - “foci”
Affiliation Networks

- Represent the affiliation of people with foci
- Affiliation networks are 2-mode networks: number of modes refers to the number of distinct kinds of social entities in a network
- Nodes of one type affiliate with nodes of the other types
- Connections among members of one of the modes are based on linkages established through the second
- Affiliation networks allow to study the dual perspectives of the actors and the events
Affiliation and Affiliation Networks

Affiliation Networks: Example

Memberships of people on corporate boards of directors:

- Two companies are implicitly linked by having the same person sit on both their boards. Higher chance for information and influence to flow between the two companies.
- Two people are implicitly linked by serving together on a board.
- Misses other contextual information: e.g. here, we see presidents of two major universities and a former US Vice-President.
Other Examples for Affiliation Networks

- Facebook.com users and groups/networks
- XING.com users and groups
- Netflix customers and movies
- Amazon customers and books
- Scientific network of authors and articles
- etc
Properties of Affiliation Networks

- Rates of Participation: the number of foci (e.g. events) with which each person is affiliated. Equal to degree of node representing the person in the bipartite graph
- Size of focus: number of people affiliated with a focus. Equal to degree of the node representing the event in the bipartite graph
- Density (i.e., mean number of events to which pairs of persons belong), reachability, connectedness, diameter (i.e., length of the longest path between any pair of persons/events) defined on one mode networks that are extracted from affiliation networks
Affiliation Network Representation

2-mode Sociomatrix:

- Rows index actors, columns events

General form: \[
\begin{pmatrix}
0 & A \\ A^T & 0
\end{pmatrix}
\]
Folding: process of transforming two mode networks into one mode networks

I.e., replacing paths of length two in the bipartite graph among actors by an (undirected) edge.

Each two mode network can be folded into 2 one mode networks

Thus, groups can be described as collections of actors affiliated with it and actors can be described as collections of groups with which they are affiliated.
Affiliation and Affiliation Networks

Example

Affiliation matrix: $X$

Co-membership matrix: $X^TX$

Event overlap matrix: $X^TX^T$
Co-Evolution of Social and Affiliation Networks

- Social networks and affiliation networks change over time: new friendship links are formed, and people become associated with new foci.
- Co-evolution that reflects the interplay between selection and social influence.
- If two people participate in a shared focus, this provides them with an opportunity to become friends.
- If two people are friends, they can influence each other’s choice of foci.
Social-Affiliation Network

- Extends notion of affiliation network - two distinct kinds of edges:
  - Edge in a social network that connects two people, and indicates friendship (or alternatively some other social relation, like professional collaboration)
  - Edge that connects a person to a focus, and indicates the participation of the person in the focus.
Suppose we have two nodes B and C with a common neighbor A in the network, and suppose that an edge forms between B and C.

- **Three types of link formation in social affiliation networks:**
  - **Triadic Closure:** If A, B, and C each represent a person, then the formation of the link between B and C is triadic closure.
  - **Focal Closure:** If B and C represent people, but A represents a focus - tendency of two people to form a link when they have a focus in common.
  - **Membership Closure:** If A and B are people, and C is a focus, then formation of a new affiliation. B takes part in a focus that her friend A is already involved in (social influence).
Example of a Social-Affiliation Network

Figure: Social-affiliation network containing both people and foci, edges can form under the effect of several different kinds of closure processes: two people with a friend in common, two people with a focus in common, or a person joining a focus that a friend is already involved in.
Example of a Social-Affiliation Network

**Figure:** Social-affiliation network containing both people and foci, edges can form under the effect of several different kinds of closure processes: two people with a friend in common, two people with a focus in common, or a person joining a focus that a friend is already involved in.
Closure Types and Social Influence vs Selection

Triadic closure

Focal closure

Membership closure

Social influence
Some thoughts on Link Formation

- Described methods good examples for small group settings
- Hard to analyse on a larger scale: track mechanisms as they operate in large populations
- Accumulation of many small effects can produce something observable in the aggregate
- In reality: many link formation processes are “off the record”
- Not easy to choose group of people (and social foci), and accurately quantify relative contributions that different mechanisms make to formation of real network links
- Approach: Through online studies, caveat - 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
Triadic Closure

- How much more likely is that edge forms between two nodes which are already friends? (triadic closure)
- Is it even more likely if nodes have more than 1 friend in common?

Figure: Anna and Esther have two friends in common, while Claire and Daniel only have one friend in common
Triadic Closure

- How much more likely is that edge forms between two nodes which are already friends? (triadic closure)
- Is it even more likely if nodes have more than 1 friend in common?

Figure: Anna and Esther have 2 friends in common, Claire and Daniel have 1 common friend

Yes! Two have more common friends, more sources of opportunity and trust for the interaction, more people with incentive to bring them together.
Empirical Assessment of Triadic Closure [Kossinets and Watts, 20066] (1/2)

We address this triadic closure empirically using network data as follows:

- Take two snapshots of the network at different times.
- For each k, we 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.
- Define $T(k)$ to be the fraction of these pairs that have formed an edge by the time of the second snapshot.
  - This is the empirical estimate for the probability that a link will form between two people with k friends in common.
  - It reflects the power of triadic closure.
- Plot $T(k)$ as a function of k to illustrate the effect of common friends on the formation of links.
Empirical Assessment of Triadic Closure [Kossinets and Watts, 20066] (2/2)

- Dataset: history of e-mail communication among 22,000 undergraduate and graduate students over a one-year period at a large US university.
- From the communication traces, Kossinets and Watts constructed a network that evolved over time, joining two people by a link at a given instant if they had exchanged e-mail in each direction at some point in the past 60 days.
- Calculation of $T(k)$: Determination of “average” version of $T(k)$ by taking multiple pairs of snapshots.
  - Building a curve for $T(k)$ on each pair of snapshots, and then averaging all obtained curves.
  - Observations in each snapshot were one day apart, so computation gives average probability that two people form a link per day, as a function of the number of common friends they have.
Results

- Black: curve determined from the data. Dotted: comparison to probabilities computed according to baseline (common friends provide independent probabilities of link formation)

- Clear evidence for triadic closure:
  - Probability of link formation increases steadily as number of common friends increases
  - Roughly linear increase, with 2 common friends more than twice the effect on link formation!
Focal Closure

- Using the same approach, (Kossinets and Watts, 2006) compute probabilities for the focal closure
- What is the probability that two people form a link as a function of the number of foci they are jointly affiliated with?
- Example: Combining university e-mail dataset with information about the class schedules for each student
- Each class became a focus, and two students shared a focus if they had taken a class together
Empirical Results

- Single shared class has similar effect on link formation as a single shared friend.
- But: curve for focal closure behaves quite differently from triadic closure curve.
  - Turns downward and appears to approximately level off, rather than turning slightly upward.
  - Thus, subsequent shared classes after the first produce a 'diminishing returns' effect.
- Interesting open question to understand how this generalizes to other types of shared foci, and to other domains.
Membership Closure

- What is the probability that a person becomes involved with a particular focus as a function of the number of friends (k) who are already involved in it?
- (Kossinets and Watts, 2006) did not provide any analysis on this
- Therefore, a study on membership closure in Wikipedia is presented here
- Two types of nodes: Wiki editors $U$ and Wiki articles $V$, link between $u \in U$ and $v \in V$ whenever $u$ and $v$ have communicated with one writing on the user talk page of the other
- Association to focus: edited the article
Observations

- Probabilities increase with the number $k$ of common neighbors—representing friends associated with the foci.
- Initial increasing effect similar as with triadic closure: the probability of editing a Wikipedia article more than twice as great when you have 2 connections into the focus rather than 1.
Question: is the homophily arising because:

- Editors are forming connections with those who have edited the same articles they have - **selection**?
- Editors are led to the articles of those they talk to - **social influence**?
- Similarity = \( \frac{\text{number of articles edited by both A and B}}{\text{number of articles edited by at least one of A or B}} \)
- Gives us neighborhood overlap of two editors
For each pair of editors A and B who have ever communicated, record their similarity over time (actors in Wikipedia are timestamped).

Time moves in discrete units, advancing by one “tick” whenever either A or B performs an action on Wikipedia.

Next, declare time 0 for the pair A and B to be the point at which they first communicated.

Results show similarity as function of time, one for each pair of editors who ever communicated.
Results

- Similarity increases both before and after moment of first interaction (both selection and social influence)
- More steep before 0 (particular role on similarity)
- Rise in similarity before two editors meet
- Note that similarity much higher for interacting users than for pairs of editors who have not interacted
- Blue line shows similarity of a random pair (non-interacting)
Homophily and Segregation

Homophily can explain segregation e.g. formation of ethnically and racially homogeneous neighborhoods in cities.

**Figure**: Map by Moebius and Rosenblatt: percentage of African-Americans per city block in Chicago in 1940 and 1960. In blocks colored yellow and orange, percentage is below 25, in brown and black above 75.
A Model for Spatial Segregation - The Schelling Model
The Schelling Model

- Thomas Crombie Schelling (born 14 April 1921)
- An American economist, and Professor of foreign affairs, national security, nuclear strategy, and arms control at the School of Public Policy at University of Maryland, College Park.
- Awarded the 2005 Nobel Memorial Prize in Economic Sciences (shared with Robert Aumann) for “having enhanced our understanding of conflict and cooperation through game-theory analysis”
The Schelling Model

- Schelling showed that a small preference for one’s neighbors to be of the same color could lead to total segregation.
- Shows how global patterns of spatial segregation can arise from homophily at local level.
- Focuses on intentionally simplified mechanism to show how forces that lead to segregation are highly robust - they can operate even when no one explicitly aims for segregated outcome.
How does the Schelling Model work? (1/2)

- Assume a population of individuals (aka agents) of type $X$ or $O$
- Types represent immutable characteristics (e.g., age)
- Two populations of the two agent types are initially placed into random locations of a neighborhood represented by a grid
- After placing all agents, each cell either occupied by an agent or empty
- The neighbor relationships among the cells can be represented very simply as a graph: cells are the nodes, edges are inserted between two cells that are neighbors on the grid

(a) Agents occupying cells on a grid.

(b) Neighbor relations as a graph.
How does the Schelling Model work? (2/2)

- Now, determine if each agent is satisfied with its current location.
- Agent is satisfied if is surrounded by at least \( t \) percent of its own type of neighboring agents.
- Threshold \( t \) applies to all agents in the model (in reality everyone might have a different threshold they are satisfied with).
- The higher \( t \), the higher the likelihood that agents will not be satisfied with their current location.
- Example:
  - For example, if \( t = 3 \), agent \( X \) is satisfied if at least 3 of its neighbors are also \( X \).
  - If fewer than 3 are \( X \), then the agent is not satisfied, and it will want to change its location in the grid.
  - When an agent is not satisfied, it can be moved to any vacant location in the grid.
  - Any algorithm can be used to choose new location (e.g., random selection, nearest available location, 1 row at a time).
Example

(a) Initial stage

(b) After one round

Figure: Left image: all dissatisfied agents have an asterisk next to them. Right image: shows new configuration after all dissatisfied agents have been moved to unoccupied cells (1 row at a time) where they are satisfied. May cause other agents to become unsatisfied, then new round of movement begins.
Netlogo Example of the Schelling Model

- http://ccl.northwestern.edu/netlogo/
- Go to File/Model Library/Social Science/Segregation
Observations from Schelling’s Model

- Spatial segregation takes place even though no individual agent actively seeks it.
- Segregation doesn’t happen due to built-in model agents that are willing to be in the minority.
- Ideally, all agents are carefully arranged in an integrated pattern.
- However, from random start hard for agents to find such integrated patterns.
- At more general level, Schelling model is an example of how fixed characteristics (e.g., ethnicity) can become highly correlated with mutable characteristics.
- E.g. decision where to live, which over time conforms to similarities in agents immutable types, producing segregation.
Summary and Practical Examples
Summary

We have learned about:

- Homophily
- Affiliation, Affiliation Networks, Social Affiliation Networks
- Link Formation in Social Affiliation Networks, 3 types of closure
- The Schelling Model as spatial model for segregation
Targeted marketing:

- User recommendations in group settings. Authors found that social ties are indicative of similar preferences. \(^1\)
- Exploited this to detect communities of users with similar taste (homophily) and to exploit this to improve Collaborative Filtering (CF)

Investigating effects in society, e.g., the glass ceiling effect \(^2\)

Investigating political communications in social networks \(^3\)

---

\(^1\) Narasimhan, et al., 2014. Exposing commercial value in social networks: matching online communities and businesses


Thanks for your attention - Questions?

elisabeth.lex@tugraz.at

Slides use figures from Chapter 4 of Networks, Crowds and Markets by Easley and Kleinberg (2010)