Social Network Analysis: homophily

Donglei Du
(ddu@unb.ca)

Faculty of Business Administration, University of New Brunswick, NB Canada Fredericton
E3B 9Y2

Table of contents

1. Homophily
   - the Schelling model
2. Test of Homophily
3. Mechanisms Underlying Homophily: Selection and Social Influence
4. Affiliation Networks and link formation
   - Affiliation Networks
   - Three types of link formation in social affiliation Network
5. Empirical evidence
   - Triadic closure: Empirical evidence
   - Focal Closure: empirical evidence
   - Membership closure: empirical evidence (Backstrom et al., 2006; Crandall et al., 2008)
   - Quantifying the Interplay Between Selection and Social Influence: empirical evidence—a case study
6. Appendix A: Assortative mixing in general
   - Quantify assortative mixing

Homophily: introduction

- The material is adopted from Chapter 4 of (Easley and Kleinberg, 2010).
- "love of the same"; "birds of a feather flock together"
- At the aggregate level, links in a social network tend to connect people who are similar to one another in dimensions like
  - Immutable characteristics: racial and ethnic groups; ages; etc.
  - Mutable characteristics: places living, occupations, levels of affluence, and interests, beliefs, and opinions; etc.
- A.k.a., assortative mixing

Homophily at action: racial segregation

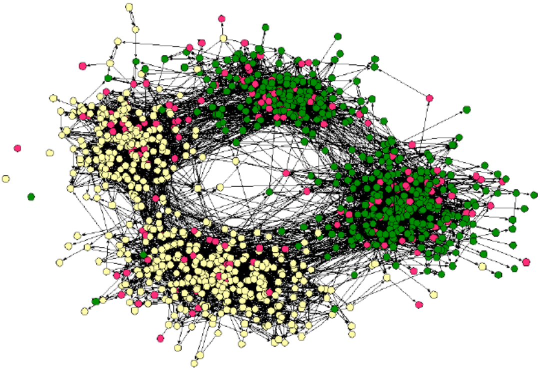
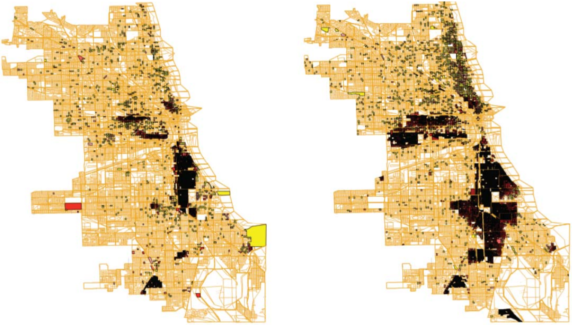


Figure 4.1: Homophily can produce a division of a social network into densely-connected, homogeneous parts that are weakly connected to each other. In this social network from a town's middle school and high school, two such divisions in the network are apparent: one based on race (with students of different races drawn as differently colored circles), and the other based on friendships in the middle and high schools respectively [304].

Figure: Credit: (Easley and Kleinberg, 2010)

Homophily at action: racial segregation Credit: (Easley and Kleinberg, 2010)



(a) *Chicago, 1940*

(b) *Chicago, 1960*

Figure 4.14: The tendency of people to live in racially homogeneous neighborhoods produces spatial patterns of segregation that are apparent both in everyday life and when superimposed on a map — as here, in these maps of Chicago from 1940 and 1960 [302]. In blocks colored yellow and orange the percentage of African-Americans is below 25, while in blocks colored brown and black the percentage is above 75.

Figure: Credit: (Easley and Kleinberg, 2010)

Homophily: 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"'.

How the Schelling model works?

- There is a population of individuals of two types.
- Each individual wants to have at least t other agents of its own type as neighbors.
- Unsatisfied individuals move in a sequence of rounds as follows
    - in each round, in a given order, each unsatisfied moves to an unoccupied cell where it will be satisfied (details can differ with similar qualitative behaviour).
    - These new locations may cause different individuals to be unsatisfied, leading to a new round of movement.

Illustration in **Netlogo**

- Social segregation by Thomas Schelling
    - Illustration in **Netlogo**
        - http://ccl.northwestern.edu/netlogo/
        - Go to File/Model Library/Social Science/Segregation

Observation from Shelling's model

- Spatial segregation is taking place even though no individual agent is actively seeking it.
- Segregation is not happening because we have subtly built it into the model–agents are willing to be in the minority, and they could all be satisfied if we were only able to carefully arrange them in an integrated pattern.
- The problem is that from a random start, it is very hard for the collection of agents to find such integrated patterns

Observation from Shelling's model

- Viewed at a still more general level, the Schelling model is an example of how characteristics that are fixed and unchanging (such as race or ethnicity) can become highly correlated with other characteristics that are mutable.
    - In this case, the mutable characteristic is the decision about where to live, which over time conforms to similarities in the agents' (immutable) types, producing segregation.

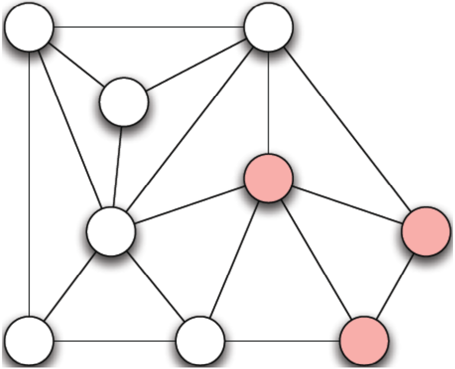
Test of Homophily

- Consider a random network $R = (V, E^r)$ where each node is assigned male with probability p, and female with probability $1 - p$.
- Let $G = (V, E)$ be a random sample of R with p fraction of male, and $1 - p$ 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, and $X_{ij} = 0$ otherwise.
  - Then X_{ij} is a Bernoulli random variable such that $P(X_{ij} = 1) = 2p(1 - p)$.
  - $\mu_0 = \mathbb{E}[X_{ij}] = 2p(1 - p)$;
    $\sigma_0^2 = \mathbb{E}[X_{ij}] = 2p(1 - p)(1 - 2p(1 - p))$.

Test of Homophily

- We can statistically test for homophily according to gender as follows: Let μ be the fraction of cross-gender edges in the random network R.
  `Homophily Test:` $H_0 : \mu \geq \mu_0$ `against` $H_1 : \mu < \mu_0$.
- If the fraction of cross-gender edges is significantly less than $2p(1-p)$, then there is evidence for homophily.

Homophily test: example



- In the above network, $p = 2/3$ and the fraction of cross-gender edges is $5/18$. On the other hand $\mu_0 = 2p(1-p) = 4/9 = 8/18$.
- Note that 5/18<8/18 showing evidence of homophily. But is this statistically significant?
- A more sensible approach is to perform a statistical hypothesis testing...

Homophily statistical hypothesis testing based on t-test: example

- We have a sample of size 18 in the example network:
  - 5 cross-edges and 13 non-cross-edges

$$111110000000000000$$

- $\mu_0 = 2p(1-p) = 4/9 \approx 0.44444$
- $H_0 : \mu \geq \mu_0$ against $H_1 : \mu < \mu_0$.
  - This is a one-tailed test.
- Due to the high interdependence of edges in networks, the standard t-test is not a good choice.

The t-test output

```
rm(list = ls())  # clear memory
# According to classical t-test x contains the 5 cross-edges and 13
# non-cross-edges
x <- c(1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
# p nad mu in the binominal distribution
p <- 2/3  #each node is assinged male with prob. p
m <- 2 * p * (1 - p)  # fraction of cross-geneder edges
# one-tailed test test H_0: mu\ge m vs H_1: mu<m
t.test(x, alternative = "less", mu = m, conf.level = 0.95)

##
##  One Sample t-test
##
## data:  x
## t = -1.534, df = 17, p-value = 0.07169
## alternative hypothesis: true mean is less than 0.4444
## 95 percent confidence interval:
##    -Inf 0.4668
## sample estimates:
## mean of x
##    0.2778
```

- Note that the sample mean $5/18 \approx 0.2777778$ is inside the 95%-CI: (-Inf, 0.4667556) based on t-test, implying that these is no strong evidence of homophily.
- But the Normal assumption in t-test is rarely satisfied fro network data.

A more sensible method: non-parametric bootstrap technique

```
library(boot)
samplemean <- function(x, d) {
    return(mean(x[d]))
}
# obtain bootstrap estimates of the standard deviation of the distribution
# of the mean:
set.seed(1)  # set seed so result can be repeated
b <- boot(x, samplemean, R = 1000)  # 1000 replications
ci <- boot.ci(b, type = "basic")  # The bootstrap CI
ci

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = b, type = "basic")
##
## Intervals :
## Level      Basic
## 95%   ( 0.0556,  0.4444 )
## Calculations and Intervals on Original Scale
```

- Note that the sample mean $5/18 \approx 0.2777778$ is inside the 95%-CI: (0.0556, 0.4444) based on bootstrap test, implying also that these is no strong evidence of homophily.

Homophily: selection vs social influence

- There are two reversing mechanisms by which ties among similar people are preferentially formed:
    - Selection: the tendency of people to form friendships with others with similar characteristics.
    - Social influence: people may modify their behaviors to bring them more closely into alignment with the behaviors of their friends.
- With selection, the individual characteristics drive the formation of links, while with social influence, the existing links in the network serve to shape people's (mutable) characteristics.

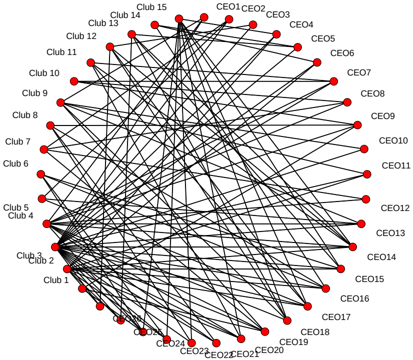
Important questions about homophily

- An observation of homophily is often not an endpoint in itself, but rather the starting point for deeper questions, such as
  - why the homophily is present;
  - how its underlying mechanisms will affect the further evolution of the network;
  - how these mechanisms interact with possible outside attempts to influence the behavior of people in the network.

How to distinguish between election and social influence: longitudinal studies

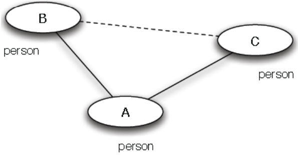
- Longitudinal studies of a social network by tracking the coevolution of the behavioral changes and the connection changes over a period of time:
  - behavioral changes after connection changes or
  - connection changes after behavioral changes.
- Moreover, we can also quantify the relative impact of these different factors:
  - how these two effects interact, and whether one is more strongly at work than the other?
- Understanding the tension between these different forces can be important not just for identifying underlying causes, but also for reasoning about the effect of possible interventions one might attempt in the system

Affiliation Networks: a.k.a. two-mode network

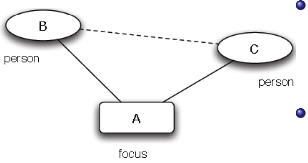


- Affiliation networks are examples of a class of graphs called bipartite graphs.
  - Namely, nodes can be divided into two sets in such a way that every edge connects a node in one set to a node in the other set. (In other words, there are no edges joining a pair of nodes that belong to the same set; all edges go between the two sets.)
- Affiliation networks represents the participation of a set of people in a set of foci.

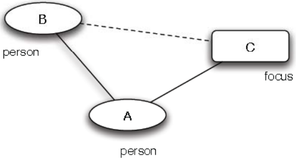
Three types of link formation in social affiliation Networks



(a) *Triadic closure*



(b) *Focal closure*



(c) *Membership closure*

- **Triadic closure**: A, B, and C are persons, the formation of the link between B and C.
- **Focal closure** (selection): B and C are people, but A is a focus.
- **Membership closure** (social influence): A and B are people, and C is a focus, then we have the formation of a new affiliation: B takes part in a focus that her friend A is already involved.

Triadic closure: Empirical evidence—a case study (Kossinets and Watts, 2006)

- Questions to be answer: “How much more likely is an edge to form between two people if they have multiple friends in common?"
- 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.

Details of Kossinets and Watts case study

- Dataset:
    - Retrieve the full history of e-mail communication among roughly 22,000 undergraduate and graduate students over a one-year period at a large U.S. 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.
- $T(k)$ calculation:
    - They then determined an “average" version of by taking multiple pairs of snapshots:
        - they built a curve for $T(k)$ on each pair of snapshots using the procedure described above, and then averaged all the curves they obtained.
        - In particular, the observations in each snapshot were one day apart, so their computation gives the average probability that two people form a link per day, as a function of the number of common friends they have.

Observation

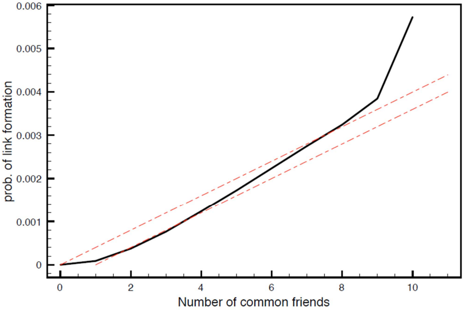


Figure: $T(k)$ is the solid black line, and the upper dashed line is the baseline model (Kossinets and Watts, 2006)

- There is clear evidence for triadic closure:
    - $T(0)$ is very close to 0, after which the probability of link formation increases steadily as the number of common friends increases.
- This probability increases in a roughly linear fashion as a function of the number of common friends, with an upward bend away from a straight-line shape.

Comparison to the baseline model

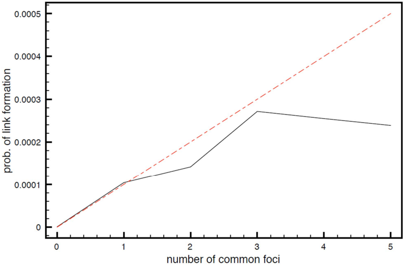
- Question to answer: “what one might have expected the data to look like in the presence of triadic closure under the assumption of independent effects from common friends?"
- For some small probability p, each common friend that two people have given them an independent probability p of forming a link each day.
- If two people have k friends in common, the probability they fail to form a link on any given day is $(1-p)^k$:
  - because each common friend fails to cause the link to form with probability $1-p$, and these k trials are independent.
- Therefore the probability that the link between two people having a common friends does form on a given day, according to the simple baseline model, is

$$T_{\text{baseline}}(k) = 1 - (1-p)^k \approx pk \text{ (for small } p\text{)}.$$

Focal Closure: empirical evidence

- 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?
- For focal closure, Kossinets and Watts supplemented their 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

Observations

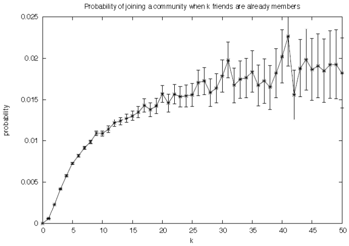


- A single shared class turns out to have roughly the same absolute effect on link formation as a single shared friend;
- But after this the curve for focal closure behaves quite differently from the curve for triadic closure:
  - it turns downward and appears to approximately level off, rather than turning slightly upward.
  - Thus, subsequent shared classes after the first produce a 'diminishing returns' effect.
- Comparing to the same kind of baseline, in which the probability of link formation with k shared classes is $1-(1-p)^k$:
  - We see that the real data turns downward more significantly than this independent model.
- Again, it is an interesting open question to understand how this effect generalizes to other types of shared foci, and to other domains.

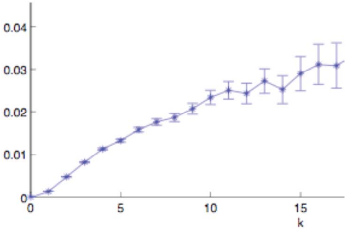
membership closure: empirical evidence

- Using the same approach, we can compute probabilities for the 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. We show the membership closure using two other case studies:
  - LiveJournal blogging (Backstrom et al., 2006):
    - Two types of nodes: users U and user-defined communities V
    - There is an edge between $u \in U$ and $v \in V$ whenever u belongs to v.
  - Wikipedia (Crandall et al., 2008)
    - Two types of nodes: Wiki editors U and Wiki articles V
    - There is an edge between $u \in U$ and $v \in V$ whenever u edits v.

Observations (Backstrom et al., 2006; Crandall et al., 2008)



LiveJournal (Backstrom et al., 2006)



Wikipedia (Crandall et al., 2008)

- The probabilities increase with the number k of common neighbors-representing friends associated with the foci.
- The marginal effect diminishes as the number of friends increases, but the effect of subsequent friends remains significant.
- Moreover, there is an initial increasing effect similar to what we saw with triadic closure: in this case, the probability of joining a LiveJournal community or editing a Wikipedia article is more than twice as great when you have two connections into the focus rather than one.
- In other words, the connection to a second person in the focus has a particularly pronounced effect, and after this the diminishing marginal effect of connections to further people takes over.

Quantifying the Interplay Between Selection and Social Influence (Crandall et al., 2008)

- Question to answer: “is the homophily arising because editors are forming connections with those who have edited the same articles they have (selection), or is it because editors are led to the articles of those they talk to (social influence)?"
- The social network
    - consist of all Wikipedia editors who maintain talk pages, and there is an edge connecting two editors if they have communicated, with one writing on the talk page of the other.
    - An editor’s behavior will correspond to the set of articles she has edited.
- One measure of similarity of editors:

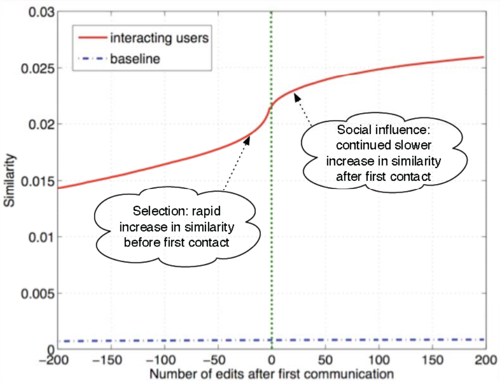
$$\frac{\text{number of articles edited by both } A \text{ and } B}{\text{number of articles edited by at least one of } A \text{ or } B}$$

    - It is precisely the neighborhood overlap of two editors in the bipartite affiliation network of editors and articles, consisting only of edges from editors to the articles they’ve edited.

Methodology

- For each pair of editors A and B who have ever communicated, record their similarity over time, where 'time' here moves in discrete units, advancing by one 'tick' whenever either A or B performs an action on Wikipedia (editing an article or communicating with another editor).
- Next, declare time 0 for the pair $A - B$ to be the point at which they first communicated.
- This results in many curves showing similarity as a function of time - one for each pair of editors who ever communicated.
- Finally, each curve is shifted so that time is measured for each one relative to the moment of first communication.
- Averaging all these curves yields a single plot in the next slide, showing the average level of similarity relative to the time of first interaction, over all pairs of editors who have ever interacted on Wikipedia.

Observations



- Similarity is clearly increasing both before and after the moment of first interaction, indicating that both selection and social influence are at work.
- However, the the curve is not symmetric around time 0:
  - the period of fastest increase in similarity is clearly occurring before 0, indicating a particular role for selection: there is an especially rapid rise in similarity, on average, just before two editors meet
- Also note that the levels of similarity depicted in the plot are much higher than for pairs of editors who have not interacted:
  - the dashed blue line at the bottom of the plot shows similarity over time for a random sample of non-interacting pairs;
  - it is both far lower and also essentially constant as time moves forward.

Future work...

- Whether the general shapes of the curves are similar across different domains
- Whether these curve shapes can be explained at a simpler level by more basic underlying social mechanisms.
- Formulate more complex, nuanced questions that can still be meaningfully addressed on large datasets.

Quantify assortative mixing

- Given a network with vertices $\{1, \ldots, n\}$, each vertex is associated an attribute x_i, discrete or continuous.
- Assortative mixing can be quantified by modularity:

$$Q = \begin{cases} \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(x_i, x_j), & \text{for discrete } x; \\ \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) x_i x_j, & \text{for continuous } x. \end{cases}$$

- Or equivalently Q can be calculated dby covariance.
- Replace each undirected edges by two opposite arcs. Let x^{in} and x_{out} be the in-degree and out-degree vectors for all arcs. Then

$$Q = \text{cov}(x^{\text{in}}, x^{\text{out}}).$$

- The interpretation is as follows:

$$Q = \begin{cases} > 0, & \text{assortative;} \\ < 0, & \text{disassortative;} \\ 0, & \text{uncorrelated.} \end{cases}$$

References I

Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. (2006). Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54. ACM.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., and Suri, S. (2008). Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 160–168. ACM.

Easley, D. and Kleinberg, J. (2010). Networks, crowds, and markets. *Cambridge Univ Press*, 6(1):6–1.

Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, 311(5757):88–90.