How Homophily Affects Learning and Diffusion in Networks

Benjamin Golub
Graduate School of Business
Stanford University

Matthew O. Jackson
Department of Economics
Stanford University

April 4, 2009
Homophily is the tendency of individuals with similar characteristics to associate with one another:
Homophily is the tendency of individuals with similar characteristics to associate with one another:
- characteristics include age, race, gender, religion, profession;
Homophily is pervasive and well-studied, but what are its effects?

Homophily is the tendency of individuals with similar characteristics to associate with one another:
- characteristics include age, race, gender, religion, profession;
- studied in sociology under that name since Lazarsfeld and Merton (1954).
Homophily is pervasive and well-studied, but what are its effects?

**Homophily** is the tendency of individuals with similar characteristics to associate with one another:
- characteristics include age, race, gender, religion, profession;
- studied in sociology under that name since Lazarsfeld and Merton (1954).

“For it often happens that some of us elders of about the same age come together and verify the old saw of like to like.” – Cephalus in Plato’s *Republic*, c. 380 BC
Homophily is pervasive and well-studied, but what are its effects?

Homophily is Strong and Pervasive

- Huge literature in sociology; documented across a variety of dimensions.
Homophily is Strong and Pervasive

- Huge literature in sociology; documented across a variety of dimensions.
  - Only 8% of Americans have anyone of another race with whom they “discuss important matters” (Marsden 1987).
Homophily is Strong and Pervasive

- Huge literature in sociology; documented across a variety of dimensions.
  - Only 8% of Americans have anyone of another race with whom they “discuss important matters” (Marsden 1987).
  - About 20% name someone of the opposite sex as their closest friend (Verbrugge 1977).
Homophily is Strong and Pervasive

- Huge literature in sociology; documented across a variety of dimensions.
  - Only 8% of Americans have anyone of another race with whom they “discuss important matters” (Marsden 1987).
  - About 20% name someone of the opposite sex as their closest friend (Verbrugge 1977).
  - In middle school, less than 10% of “expected” cross-race friendships exist (Shrum et. al. 1988).
Friendships in a High School

How Homophily Affects Learning in Networks

Currrarini, Jackson, and Pin (2009)
Motivation

Homophily is pervasive and well-studied, but what are its effects?

Model

But What are its Effects?

Results

Data

What are the actual consequences of homophily for important processes?
But What are its Effects?

- What are the actual consequences of homophily for important processes?
- In this project, we focus on communication and build models of:
But What are its Effects?

- What are the actual consequences of homophily for important processes?
- In this project, we focus on communication and build models of:
  - networks with homophily;
But What are its Effects?

- What are the actual consequences of homophily for important processes?
- In this project, we focus on communication and build models of:
  - networks with homophily;
  - diffusion or learning processes happening in them.
But What are its Effects?

- What are the actual consequences of homophily for important processes?
- In this project, we focus on communication and build models of:
  - networks with homophily;
  - diffusion or learning processes happening in them.
- Study how homophily affects the speed of the processes.
Main Results

Homophily does not affect the spread of “news” or “rumors”.
Motivation
Model
Results
Data

Homophily is pervasive and well-studied, but what are its effects?

Main Results

- Homophily does not affect the spread of “news” or “rumors”.
- But slows
Homophily is pervasive and well-studied, but what are its effects?

Main Results

- Homophily does not affect the spread of “news” or “rumors”.
- But slows
  - convergence to consensus opinions;
Homophily is pervasive and well-studied, but what are its effects?

Main Results

- Homophily does not affect the spread of “news” or “rumors”.
- But slows
  - convergence to consensus opinions;
  - convergence to equilibrium under myopic updating.
There are $n$ agents, indexed by a set $N = \{1, 2, \ldots, n\}$. 

Benjamin Golub and Matthew O. Jackson

How Homophily Affects Learning in Networks
Multi-Type Random Network

- There are \( n \) agents, indexed by a set \( N = \{1, 2, \ldots, n\} \).
- Partitioned into \( m \) types: \( N_1, N_2, \ldots, N_m \).
There are $n$ agents, indexed by a set $N = \{1, 2, \ldots, n\}$.

Partitioned into $m$ types: $N_1, N_2, \ldots, N_m$.

The probability that an agent of type $k$ has an (undirected) link to an agent of type $\ell$ is $P_{k\ell}$. 

How Homophily Affects Learning in Networks
Multi-Type Random Network

- There are $n$ agents, indexed by a set $N = \{1, 2, \ldots, n\}$.
- Partitioned into $m$ types: $N_1, N_2, \ldots, N_m$.
- The probability that an agent of type $k$ has an (undirected) link to an agent of type $\ell$ is $P_{k\ell}$.
- Links are formed independently.
Islands Model

Special case for this talk:
Islands Model

Special case for this talk:

- All types have the same size.
Islands Model

Special case for this talk:

- All types have the same size.
- Only two probabilities:

\[ P_{k\ell} = \begin{cases} 
\rho_s & \text{if } k = \ell \\
\rho_d & \text{otherwise}.
\end{cases} \]
Measuring Homophily (in the Islands Model)

- Let $p$ be the overall link density.
Measuring Homophily (in the Islands Model)

- Let $p$ be the overall link density.
- Unnormalized homophily:

$$H = \frac{p_s}{p} \in [0, m].$$
Measuring Homophily (in the Islands Model)

- Let $p$ be the overall link density.
- Unnormalized homophily:
  \[
  H = \frac{p_s}{p} \in [0, m].
  \]
- Normalized homophily:
  \[
  h = \frac{1}{m} \frac{p_s}{p} \in [0, 1].
  \]
Shortest Path Based Communication

Any process where the time for $i$ and $j$ to communicate is proportional to the distance between them.
Shortest Path Based Communication

- Any process where the time for $i$ and $j$ to communicate is proportional to the distance between them.
- Examples:
Shortest Path Based Communication

- Any process where the time for $i$ and $j$ to communicate is proportional to the distance between them.

Examples:
- Sending targeted orders through an organizational chart.
Shortest Path Based Communication

- Any process where the time for $i$ and $j$ to communicate is proportional to the distance between them.
- Examples:
  - Sending targeted orders through an organizational chart.
  - Broadcasting.
Broadcasting

Communication Process 1: Shortest Path (Diffusion)
Communication Process 2: Linear Updating (Learning)
Broadcasting

Communication Process 1: Shortest Path (Diffusion)
Communication Process 2: Linear Updating (Learning)
Broadcasting
Broadcasting

Communication Process 1: Shortest Path (Diffusion)
Communication Process 2: Linear Updating (Learning)
Broadcasting

Benjamin Golub and Matthew O. Jackson

How Homophily Affects Learning in Networks
How Homophily Affects Learning in Networks

Benjamin Golub and Matthew O. Jackson
Motivation
Model
Results
Data

Networks
Communication Process 1: Shortest Path (Diffusion)
Communication Process 2: Linear Updating (Learning)

Broadcasting

Benjamin Golub and Matthew O. Jackson
How Homophily Affects Learning in Networks
A sufficient statistic for time to communicate (in a *given, fixed* network) in this case is just the expected distance between two randomly chosen nodes.
Linear Updating (French 1956, DeGroot 1974)

The belief of agent $i$ at time $t + 1$ is an average of the beliefs of his neighbors at time $t$. 
Linear Updating (French 1956, DeGroot 1974)

The belief of agent $i$ at time $t + 1$ is an average of the beliefs of his neighbors at time $t$.

$$b_i(t + 1) = \sum_{j} \frac{A_{ij}}{d_i} b_j(t),$$
Communication Process 2: Linear Updating (Learning)

The belief of agent $i$ at time $t + 1$ is an average of the beliefs of his neighbors at time $t$.

$$b_i(t + 1) = \sum_j \frac{A_{ij}}{d_i} b_j(t),$$

where

$$A_{ij} = \begin{cases} 
1 & \text{if } i \text{ and } j \text{ are linked} \\
0 & \text{otherwise}.
\end{cases}$$
Linear Updating (French 1956, DeGroot 1974)

The belief of agent $i$ at time $t + 1$ is an average of the beliefs of his neighbors at time $t$.

$$b_i(t + 1) = \sum_j \frac{A_{ij}}{d_i} b_j(t),$$

where

$$A_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked} \\ 0 & \text{otherwise.} \end{cases}$$

and $d_i = \# \{\text{neighbors of } i\}$
The belief of agent $i$ at time $t + 1$ is an average of the beliefs of his neighbors at time $t$.

$$b_i(t + 1) = \sum_j \frac{A_{ij}}{d_i} b_j(t),$$

where

$$A_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked} \\ 0 & \text{otherwise.} \end{cases}$$

and $d_i = \#\{\text{neighbors of } i\}$
Linear Updating (French 1956, DeGroot 1974)

The belief of agent $i$ at time $t + 1$ is an average of the beliefs of his neighbors at time $t$.

$$b_i(t + 1) = \sum_j A_{ij} \frac{d_i}{b_j(t)},$$

where

$$A_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked} \\ 0 & \text{otherwise}. \end{cases}$$

and $d_i = \#\{\text{neighbors of } i\}$
The belief of agent $i$ at time $t + 1$ is an average of the beliefs of his neighbors at time $t$.

$$b_i(t + 1) = \sum_j A_{ij} \frac{1}{d_i} b_j(t),$$

where

$$A_{ij} = \begin{cases} 
1 & \text{if } i \text{ and } j \text{ are linked} \\
0 & \text{otherwise.}
\end{cases}$$

and $d_i = \#\{\text{neighbors of } i\}$

$$b_1(t + 1) = \frac{1}{2} b_1(t) +$$
Linear Updating (French 1956, DeGroot 1974)

The belief of agent $i$ at time $t + 1$ is an average of the beliefs of his neighbors at time $t$.

$$b_i(t + 1) = \sum_j A_{ij} \frac{d_i}{d_i} b_j(t),$$

where

$$A_{ij} = \begin{cases} 
1 & \text{if } i \text{ and } j \text{ are linked} \\
0 & \text{otherwise.} 
\end{cases}$$

and $d_i = \#\{\text{neighbors of } i\}$

$$b_1(t + 1) = \frac{1}{2} b_1(t) + \frac{1}{2} b_2(t)$$
Think of $b_i(t)$ as a behavior, not a belief.
Linear Updating as Myopic Best-Response

- Think of $b_i(t)$ as a \textit{behavior}, not a \textit{belief}.
- Utilities:

$$u_i(t) = -\sum_j \frac{A_{ij}}{d_i} (b_i(t) - b_j(t))^2$$

Note that everyone choosing the same behavior is an equilibrium. But which behavior? Agents best-respond to last period’s choices. This gives the linear updating process.
Think of $b_i(t)$ as a behavior, not a belief.

Utilities:

$$u_i(t) = - \sum_j \frac{A_{ij}}{d_i} (b_i(t) - b_j(t))^2$$

Note that everyone choosing the same behavior is an equilibrium.
Think of $b_i(t)$ as a *behavior*, not a *belief*.

Utilities:

$$u_i(t) = -\sum_j \frac{A_{ij}}{d_i} (b_i(t) - b_j(t))^2$$

Note that everyone choosing the same behavior is an equilibrium. *But which behavior?*
Think of $b_i(t)$ as a behavior, not a belief.

Utilities:

$$u_i(t) = -\sum_j \frac{A_{ij}}{d_i} (b_i(t) - b_j(t))^2$$

Note that everyone choosing the same behavior is an equilibrium. But which behavior?

Agents best-respond to last period’s choices.
Think of $b_i(t)$ as a *behavior*, not a *belief*.

Utilities:

$$u_i(t) = -\sum_j A_{ij} \frac{d_i}{d_i} (b_i(t) - b_j(t))^2$$

Note that everyone choosing the same behavior is an equilibrium. **But which behavior?**

Agents best-respond to last period’s choices.

This gives the linear updating process.
Linear Updating

Communication Process 1: Shortest Path (Diffusion)

Communication Process 2: Linear Updating (Learning)

$t = 0$
Linear Updating

Communication Process 1: Shortest Path (Diffusion)
Communication Process 2: Linear Updating (Learning)
How Homophily Affects Learning in Networks

Communication Process 1: Shortest Path (Diffusion)

Communication Process 2: Linear Updating (Learning)
Linear Updating

Communication Process 1: Shortest Path (Diffusion)
Communication Process 2: Linear Updating (Learning)

$t = 4$
Linear Updating

Communication Process 1: Shortest Path (Diffusion)
Communication Process 2: Linear Updating (Learning)
Measuring Speed with Linear Updating

Idea of the measure: how long does it take to get close to consensus (in a \textit{given}, \textit{fixed} network)?
Measuring Speed with Linear Updating

- Idea of the measure: how long does it take to get close to consensus (in a *given, fixed* network)?
- Requires measuring how close we are to consensus at time $t$. 

Benjamin Golub and Matthew O. Jackson

*How Homophily Affects Learning in Networks*
Measuring Speed with Linear Updating

- Idea of the measure: how long does it take to get close to consensus (in a given, fixed network)?
- Requires measuring how close we are to consensus at time $t$.
- A measure of “how close” at time $t$:
  - Consider a random opinion transmitted at time $t$. 

Benjamin Golub and Matthew O. Jackson
How Homophily Affects Learning in Networks
Measuring Speed with Linear Updating

- Idea of the measure: how long does it take to get close to consensus (in a given, fixed network)?
- Requires measuring how close we are to consensus at time $t$.
- A measure of “how close” at time $t$:
  - Consider a random opinion transmitted at time $t$.
  - What is its squared deviation from the eventual consensus?
Measuring Speed with Linear Updating

- Idea of the measure: how long does it take to get close to consensus (in a *given, fixed* network)?
- Requires measuring how close we are to consensus at time $t$.
- A measure of “how close” at time $t$:
  - Consider a random opinion transmitted at time $t$.
  - What is its squared deviation from the eventual consensus?
  - $\sqrt{\text{the expectation of that random variable}}$
  - is the distance from consensus.
Measuring Speed with Linear Updating

- Idea of the measure: how long does it take to get close to consensus (in a *given, fixed* network)?
- Requires measuring how close we are to consensus at time $t$.
- A measure of “how close” at time $t$:
  - Consider a random opinion transmitted at time $t$.
  - What is its squared deviation from the eventual consensus?
  - $\sqrt{\text{the expectation of that random variable}}$

is the distance from consensus.
(Essentially root-mean-squared distance from consensus.)
Measuring Speed with Linear Updating

Definition

The \emph{consensus time} $CT(\epsilon; A)$ is the time it takes in network $A$ until the distance from consensus remains below $\epsilon$, in the worst case, assuming beliefs start in $[0, 1]$. 
The Big Picture: How Communication Speed Depends on Density and Homophily

<table>
<thead>
<tr>
<th>Process</th>
<th>Independent Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Density</td>
</tr>
<tr>
<td>Shortest Path</td>
<td>↑</td>
</tr>
<tr>
<td>Linear Updating</td>
<td>0</td>
</tr>
</tbody>
</table>

Arrows indicate how communication speed is affected when the independent variable is increased.
An Approximation Notion

Definition

\[ f(n) \approx g(n) \]

means that for any \( \delta > 0 \),

\[
\mathbb{P} \left[ \frac{f(n)}{g(n)} \in (1/2 - \delta, 2 + \delta) \right] \xrightarrow{n \to \infty} 1.
\]
How Homophily Affects Shortest Path Based Communication: Assumptions

\[ d(n) := np(n) \geq (1 + \varepsilon) \log n \quad \text{for some } \varepsilon > 0 \]

(the network is dense enough that it is a. s. connected)
How Homophily Affects Shortest Path Based Communication: Assumptions

- \( d(n) := np(n) \geq (1 + \varepsilon) \log n \) for some \( \varepsilon > 0 \)
  (the network is dense enough that it is a.s. connected)

- \( \frac{\log d(n)}{\log n} \to 0 \)
  (network is not too close to complete)
How Homophily Affects Shortest Path Based Communication: Assumptions

- $d(n) := np(n) \geq (1 + \varepsilon) \log n$ for some $\varepsilon > 0$
  
  (the network is dense enough that it is a.s. connected)

- $\frac{\log d(n)}{\log n} \to 0$

  (network is not too close to complete)

- $h(n) \leq \bar{h}$ for some $\bar{h} < 1$

  (islands are not completely introspective)
Theorem (Jackson 2008)

Under the assumptions just stated,

$$\text{average distance} \approx \frac{\log n}{\log d(n)}$$

and, asymptotically, does not depend at all on homophily.
Density, not Homophily, Matters for Shortest Path Communication

Theorem (Jackson 2008)

Under the assumptions just stated,

\[ \text{average distance} \approx \frac{\log n}{\log d(n)} \]

and, asymptotically, does not depend at all on homophily.

- Homophily doesn’t matter.
Density, not Homophily, Matters for Shortest Path Communication

Theorem (Jackson 2008)
Under the assumptions just stated,

\[
\text{average distance} \approx \frac{\log n}{\log d(n)}
\]

and, asymptotically, does not depend at all on homophily.

- Homophily doesn’t matter.
- Only density matters (more = faster).
Density, not Homophily, Matters for Shortest Path Communication

- Density and homophily assumptions guarantee that the network is not too far from a tree.
Density, not Homophily, Matters for Shortest Path Communication

- Density and homophily assumptions guarantee that the network is not too far from a tree.
- So extended neighborhoods still expand exponentially.
Density, not Homophily, Matters for Shortest Path Communication

- Density and homophily assumptions guarantee that the network is not too far from a tree.
- So extended neighborhoods still expand exponentially.
- Thus, the average agent can still reach the same number of people after $t$ steps, with or without homophily.
Density, not Homophily, Matters for Shortest Path Communication

- Density and homophily assumptions guarantee that the network is not too far from a tree.
- So extended neighborhoods still expand exponentially.
- Thus, the average agent can still reach the same number of people after $t$ steps, with or without homophily.
  - Homophily does change who is close and who is far; the first hearers of the news are predominantly of the originator’s type.
Density and homophily assumptions guarantee that the network is not too far from a tree.

So extended neighborhoods still expand exponentially.

Thus, the average agent can still reach the same number of people after $t$ steps, with or without homophily.

Homophily does change who is close and who is far; the first hearers of the news are predominantly of the originator’s type.

But order does not matter – only the overall speed at which the news spreads.
Homophily, not Density, Matters for Linear Updating

Theorem

If \( d(n) / \log^2 n \to \infty \) and \( m \to \infty \)

\[
CT \left( \frac{\gamma}{n}; A(n) \right) \approx \frac{\log n}{\log(h^{-1})}
\]

where the network \( A(n) \) is the islands network with

- \( n \) nodes
- \( m \) islands
- homophily \( h \).
Homophily, not Density, Matters for Linear Updating

CT

\( h \)

Benjamin Golub and Matthew O. Jackson

How Homophily Affects Learning in Networks
Homophily, not Density, Matters for Linear Updating

- Homophily matters (more = slower).
Homophily, not Density, Matters for Linear Updating

- Homophily matters (more = slower).
- Beyond a low threshold, density doesn’t matter.
Basic intuition: each island reaches its own internal consensus, and if islands put low weight outside themselves, then it will take a long time for the differences to erode.
Homophily, not Density, Matters for Linear Updating

Steps of proof:
Homophily, not Density, Matters for Linear Updating

Steps of proof:

\[ b_i(t + 1) = \sum_j \frac{A_{ij}}{d_i} b_j(t) \]

can be written as

\[ b(t) = T^t b(0). \]
Homophily, not Density, Matters for Linear Updating

Steps of proof:

\[ b_i(t + 1) = \sum_j \frac{A_{ij}}{d_i} b_j(t) \]

can be written as

\[ \mathbf{b}(t) = \mathbf{T}^t \mathbf{b}(0). \]

Convergence of this process to steady state is controlled by second largest eigenvalue in magnitude of \( \mathbf{T} \).
Homophily, not Density, Matters for Linear Updating

Steps of proof (continued):
Steps of proof (continued):

- For a multi-type random network, we can look at a representative agent matrix with
Steps of proof (continued):

- For a multi-type random network, we can look at a *representative agent matrix* with
  - one agent for each type;
Steps of proof (continued):

- For a multi-type random network, we can look at a representative agent matrix with:
  - one agent for each type;
  - realized links replaced by expected link densities.
Steps of proof (continued):

- For a multi-type random network, we can look at a representative agent matrix with
  - one agent for each type;
  - realized links replaced by expected link densities.

- Theorem: the second eigenvalue of the big random matrix is well-approximated by the second eigenvalue of the small deterministic matrix.
Representative Agent Matrix
Representative Agent Matrix
Representative Agent Matrix
Representative Agent Matrix
The Data

- Adolescent Health data set.
- 84 schools (2 outliers removed).
- For each student:
  - grade in school (6–12);
  - gender;
  - race.
- Friendships.
Testing the Shortest Path Theorem

Recall that the theorem predicts

$$\text{average distance} \approx \frac{\log n}{\log d(n)}.$$
Testing the Shortest Path Theorem

Average Shortest Path vs \( \log(n)/\log(d) \)

without homophily: \( R^2 = 0.93 \)

Benjamin Golub and Matthew O. Jackson

How Homophily Affects Learning in Networks
Testing the Shortest Path Theorem

Average Shortest Path vs Log(n)/Log(d)

without homophily: $R^2 = 0.93$

with homophily: $R^2 = 0.94$
Recall that the theorem predicts

\[ \text{CT} \left( \gamma/n; A(n) \right) \approx \frac{\log n}{\log(h^{-1})}. \]
Testing the Consensus Time Theorem

- Recall that the theorem predicts

\[ CT(\gamma/n; A(n)) \approx \frac{\log n}{\log(h^{-1})}. \]

- Slightly fancier: replace \( h \) by \( \frac{H-1}{m-1} \), where \( H = \frac{p_s}{p_d} \) and \( m \) is number of islands.
Recall that the theorem predicts

$$\text{CT}(\gamma/n; A(n)) \approx \frac{\log n}{\log(h^{-1})}.$$  

Slightly fancier: replace $h$ by $\frac{H - 1}{m - 1}$, where $H = \frac{ps}{pd}$ and $m$ is number of islands.

Can manipulate this around and find a function $\rho$ so that

$$\rho(\text{CT}) - c \propto \frac{H - 1}{m - 1}.$$
Testing the Consensus Time Theorem

\[ R^2 = 0.231 \]