Content and Causality in Influence Networks

ICWSM
July 20, 2011

@sinanaral
Dear Sinan,

We would like to invite you to give a plenary keynote talk for 50-to-60 minutes at the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011). ICWSM 2011 will be held on July 17-20, 2011 in Barcelona (Spain) and will be collocated with IJCAI 2011. For more details see http://www.icwsm.org.

Sinan Aral wrote:

    pff - are you kidding? I'm in.

    :)!!

*Sinan Aral*
NYU, Stern School of Business
Affiliated Faculty, MIT
http://pages.stern.nyu.edu/~sara
BUY COMMERCES GETS SOCIAL

HOW YOUR NETWORKS ARE DRIVING WHAT YOU BUY

FEATURED ON YOUR FAVOURITE ONLINE SOCIAL NETWORKS
FACEBOOK TWITTER YOUTUBE BLIPPY FOURSQUARE AMazon linkedin GO TRY IT ON shopkick Groupon
“This Revolution Will Be Tweeted.”
CLICKING 🎉, YOU CAN SEE THE REPORTS THAT ARE POSTED IN THE AREA.
Causality

Content
Causality

Content
Influence - Current Formalizations

(e.g. in Studies of Twitter)

1. Influence = In Degree (Number of Followers)
2. Influence = f(Network Position) (e.g. Centrality, Betweenness)
3. Influence = Volume of Information Broadcast (Number of Tweets)
4. Influence = Breadth of Information Broadcasts (Tweets*Followers)
5. Influence = Breadth and Novelty of Information Broadcasts (Novel Tweets*Followers)
6. Influence = Breadth of Information Cascades (Tweets*Followers + Retweets*Followers of Followers and so on)
7. Influence = Triggers of Broad Novel Information Cascades (Novel Tweets*Followers + Novel Retweets*Followers of Followers)
8. Influence = Peer Behavior Adoption at time t + 1
Social Influence: how the behaviors of one’s peers change the likelihood that (or extent to which) one engages in a behavior.

Implications of this Definition

1. **Social Influence is about causal behavior change:**
   How peer behaviors change the likelihood that or extent to which one will engage in the behavior.

2. **Room for multiple social processes:**
   Change in Utility Function or Change in Perception of the Behavior or Product... Persuasion or Awareness; Imitation or Social Learning.

3. **Enables Consideration of Systems of Behavior**
   Don’t need activation on the focal behavior in question. Contrasts assumptions in marketing and innovation diffusion literature. Enables considerations of complementary behaviors.

4. **Markovian assumptions can be strict or relaxed.**
   Should influence processes be memoryless or favor recency? Cumulative effects vs. enthusiasm or excitement effects.
I used to think correlation implied causation.

Then I took a statistics class. Now I don't.

Sounds like the class helped.

Well, maybe.
Reflection and Endogeneity

Manski (1993)

Identifying **Causal** Peer Effects in Networks is Notoriously Difficult

Now lots of empirical evidence that **human behaviors** tend to **cluster** in **network space and time**, ...

**but is this because of peer influence or alternate explanations?**
“Obesity is Contagious”

Christakis & Fowler (2007)
Homophily

“Birds of a feather, flock together”

Robert Burton (1577-1640)
“Everything is Contagious”

Alternative explanations:

- Reflection
- Homophily
- Confounding

Causal Structure of the underlying dynamic process implies:

1. **Different Diffusion Properties** and
2. **Different Optimal Containment or Promotion Policies**.
Causal Estimation in Networks

   - Use variation in group size or structure as instrumental variables to identify deviations from group means.

2. **Actor Oriented Models (Co-Evolution of Networks and Behavior)** (e.g. Snijders 2001, Steglich, Snijders and Pearson 2004, Aral 2010)
   - Model micro decisions that maximize Behavioral and Network utility.
   - Apply Continuous Time Markov Models to Panel Network Data.
   - Estimate with MCMC or other simulated method of moments techniques

3. **Natural Experiments, Instrumental Variables** (Sacredote 2001, Tucker 2008)

4. **Structural Models** (Ghose and Han 2010)

5. **Automated Discovery of QED** (Jensen et al 2008)

   - Treatment – those with \( n \) friends who adopted at or before time \( t \).
   - Control – those as likely to have \( n \) friends ...”...as the treated, but who do not.

7. **Randomized Trials in Massive Networks** (Aral & Walker 2010a, b; Aral 2011)
Dynamic Matched Sampling

- Global IM Network of 27 Million Users from Yahoo! (Daily Traffic)
- Detailed demographics and geographic data.
- Comprehensive, detailed and precise data on online behaviors/activities.
- Day by Day adoption and usage of a mobile service application (Yahoo Go) launched in July 2007. (532,365 Adopters)

Distinguishing Influence from Homophily

“Influence” Estimates Comparing Adoption in Treated and Untreated Cases Under **Randomized Matching** Over Time (Methods used by those who take AM as evidence of influence)

“Influence” Estimates Comparing Adoption in Treated and Untreated Cases In Our **Dynamic Matched Sampling Framework** Over Time

Much of the estimated influence is really observable homophily.
Exaggerated Homophily Amongst Early Adopters

Cosine Distances Of Vectors of Observable Demographic, Geographic and Behavioral Data

Graph showing the cosine similarity over time for adopter friends, non-adopter friends, and random users.
Can we engineer products so they are more likely to be viral shared?

Viral Product Design

The process of explicitly engineering products so they are more likely to be shared amongst peers.

Viral Product Characteristics

Content likely to inspire viral sharing


Viral Product Features

Modalities of use

Invites
Notifications
Hypertext Embedding
(No literature)
Viral Feature Space

PERSONALIZATION INCREASING

BROADCAST

PERSONALIZED

Population

Peer Network

Specific Individuals

ACTIVITY INCREASING

ACTIVE

High Effort

Minimal Effort

Passive

No Effort

Generalized Hypertext Embedding

Personalized Hypertext Embedding

Automated Broadcast Notifications

Personalized Referrals

Collaborative Bookmarking

Automated Targeted Notifications

Greater marginal peer influence per message

More messages generated

No Effort

Minimal Effort

High Effort
Randomly Enabled Viral Messaging.
Observed the Adoption and Use of the App by Friends of Control and Experimental Group Users.
Data

- ~ 10K Experimental Users
- ~ 1.4M Friends of Experimental Users
- We Observe Application Diffusion Over this Network
  - Adoption
  - Use
Flixster - An Example Facebook Application

Featured DVD: Own it 12/8 on Blu-Ray™ & DVD

Public Enemies
If you love Johnny Depp, don’t miss Public Enemies!
65% liked it
Johnny Depp, Christian Bale, Marion Cotillard, Billy Crudup, Jason Clarke, David Wenham, Christian Stolte, Stephen Dorff
From award-winning director Michael Mann (Heat, Collateral) comes the film inspired by one of the country's most captivating and infamous outlaws John Dillinger.
Johnny Depp (Pirates of the Caribbean series) stars as the charismatic and elusive... (read more)

Want to see it: Jenna
You: WANT TO SEE IT NOT INTERESTED

Opening This Week

Armored (Armoured)
66% want to see it
Matt Dillon, Jean Reno, Laurence Fishburne, Skeet Ulrich, Amaury Nolasco, Columbus Short
A crew of officers at an armored transport security firm risk their lives...
Flixster - An Example Facebook Application

Flixster Watch movies. Tell friends.

Recent Comments

None of your friends have made any comments yet. Maybe you need more friends. Invite some friends. Or share some of your reviews.

Friends' Recent Reviews

What? Your friends haven't written any reviews? Invite some more friends and reviews will start to show up.

My Friends (81)

Movies Friends Want To See

Jenna wants to see:
The Informant!
Old Dogs
The Blind Side
Users can invite their friends to adopt the application and join their social network on the application itself.
Users can invite friends to install this application and include a personal message. Users can invite their friends to adopt the application and join their social network on the application itself.
Users receive notifications in a designated area.

When a user clicks on the referral link in the notification, they are taken to the application canvas page.
The application canvas page provides prospective users with details about the nature of the application and gives them the option to install the application.
<table>
<thead>
<tr>
<th>Baseline Group</th>
<th>Passive-broadcast Group</th>
<th>Active-personalized Group</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Diagram" /></td>
<td><img src="image2.png" alt="Diagram" /></td>
<td><img src="image3.png" alt="Diagram" /></td>
</tr>
</tbody>
</table>

- **Friends of baseline group**
- **Friends of passive-broadcast group**
- **Friends of active-personalized group**
Leakage and contamination could occur if peers are

a) connected through indirect pathways,

b) connected to multiple treated peers in different treatment groups or

c) connected to multiple treated peers in the same treatment group.

These scenarios are rare in our data

We control for leakage and peers of multiple treated users by only evaluating recruited users and right censoring contaminated peers.

Contaminated: Any peer with multiple treated peers after time $t$ at which they have multiple treated peers.

This may make our results more conservative but also minimizes leakage and contamination.
Conventional Approach in Observational Data

\[ P(y_{it} = 1 \mid y_{it-1} = 0) = F(x_{it}, \beta \sum_j w_{ij} y_{jt}) \]

“Inside-Out” Estimation

A Variance Corrected Stratified Proportional Hazards Model

\[ \lambda_k(t, X_{ki}) = \lambda_{0k}(t) e^{X_{ki} \beta} \]
## Which Features Spread Contagion Best?

<table>
<thead>
<tr>
<th></th>
<th>Personal Invitations</th>
<th>Passive Awareness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Influence Per Message</strong></td>
<td>↑ 6%</td>
<td>↑ 2%</td>
</tr>
<tr>
<td><strong>Global Diffusion</strong></td>
<td>↑ 98%</td>
<td>↑ 246%</td>
</tr>
<tr>
<td><strong>Stickiness</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sustained Engagement

Average Activity Over Time

Average Activity Per User vs days after install
<table>
<thead>
<tr>
<th>Feature</th>
<th>Personal Invitations</th>
<th>Passive Awareness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influence Per Message</td>
<td>↑ 6%</td>
<td>↑ 2%</td>
</tr>
<tr>
<td>Global Diffusion</td>
<td>↑ 98%</td>
<td>↑ 246%</td>
</tr>
<tr>
<td>Stickiness</td>
<td>↑ 17%</td>
<td>0%</td>
</tr>
</tbody>
</table>
Peer Adoption

Engagement
Causality

Content
Causality

Content
Content Gives…

Meaning
Intension
Relevance
Persuasion
Reaction
Context
Some Examples of Research


- And tons of really interesting research at ICWSM!
Opportunities for “Brokerage”… are typically enabled by weak ties

Bridging Weak Ties span “Structural Holes”
Information Advantage

Value of information comes from its uneven distribution across local network neighborhoods.

Connection to diverse neighborhoods gives access to novel pools of information.

Novel information is valuable due to its local scarcity.

Actors with scarce, novel information can
✓ broker opportunities, engage in information arbitrage
✓ use information as a commodity, or
✓ apply information to problems that are intractable given local information (innovation).
A 40 year old assumption…

- Network structure is associated with performance.
  - Productivity of R&D teams (Reagans & Zuckerman 2001)
  - Wages, Promotion, Job Placement (Burt 1992)
  - Innovation (Burt 2004)
  - Productivity of information workers (Aral et al 2006)

- Key theoretical mechanism: access to information.
The Theory is Problematic because Structural Diversity is likely to be associated with weak ties, creating A Tradeoff Between Network Diversity and Channel Bandwidth.

A Constrained Network of Strong, High Bandwidth Ties

A Diverse Network of Weak, Low Bandwidth Ties

Diversity ↔ weak ties => lower bandwidth, frequency, topical dimension, detail, complexity.

1. **Information Overlap** – The degree to which topics are uniformly or heterogeneously distributed over nodes.

2. **Size of the Topic Space** – How many distinct topics exist in the network.

3. **Refresh Rate** – How often actors’ information is changing per unit time.
### Study Context - Executive Search

![Firm Communication Network](image)

#### Communication Network: Firm Headquarters

<table>
<thead>
<tr>
<th></th>
<th>FIRM</th>
<th>HQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recruiters</td>
<td>73</td>
<td>34</td>
</tr>
<tr>
<td>Average Density</td>
<td>5.41 (19.08)</td>
<td>11.02 (32.44)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Network Constraint</th>
<th>Information Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partner</td>
<td>.24 (.14)</td>
<td>.59 (.12)</td>
</tr>
<tr>
<td>Consultant</td>
<td>.30 (.18)</td>
<td>.55 (.14)</td>
</tr>
<tr>
<td>Researcher</td>
<td>.32 (.20)</td>
<td>.55 (.15)</td>
</tr>
<tr>
<td>Mean</td>
<td>.29 (.17)</td>
<td>.57 (.14)</td>
</tr>
</tbody>
</table>
Measuring Information Diversity

Vector Space Model
Represent each email as a multidimensional vector of term frequencies
Represent In-boxes and Out-boxes as collections of email vectors
Measure ‘variance’ of the vectors in someone’s in-box or out-box
E.g. Variance of cosine distances between email vectors and information theoretic measures

Validate diversity measurement using public data from Wikipedia.org
Results

1. Diversity and Bandwidth **Tradeoff**… And both predict access to novel information.

2. Whether Brokers receive more novel information per unit time depends on the **information environment in which the broker is situated**.
   
a) Greater Information Overlap – Bandwidth more valuable than diversity.

b) Larger Topic Space – Bandwidth more valuable than diversity.

c) Faster the Refresh Rate – Bandwidth more valuable than diversity.
1. Strength of Weak Ties and Structural Hole Theory is *context dependent*.

2. Wide Bridges vs. Thick Bridges – which is more important for Complex Contagions (Centola and Macy 2007)?

- Wide Bridge – many ties – reinforcement from multiple parties.
- Thick Bridges – few strong ties – reinforcement from a *trusted* party with *detailed and comprehensive* information about the behavior.
- Which is more important? Under what conditions?
- More research is needed.
“The Cramer Effect”


LDA on of show transcripts – was Cramer “more persuasive” when he was making particular arguments.

How prior knowledge and attention impacted his “influence”

Collected Google Search, News, Trading history data for each stock in the two weeks before and after each recommendation (and ran LDA on the news as well).

Found:

1. A Selection Effect – Cramer chooses to recommend stocks trending in attention (news, search, trade volume) *just prior* to his recommendation.

2. Importance of Novelty – Boosts his influence *and* this effect more important for some arguments than others.
Causality

Content
Collaborators:

Erik Brynjolfsson, Marshall Van Alstyne, Panos Ipeirotis, Arun Sundararajan, Lev Muchnik, Dylan Walker, Sean Taylor

Generous Funding and Data:

NSF, Microsoft, Yahoo, IBM, Facebook, Cisco, SAP, France Telecom, the Marketing Science Institute, the Institute for Information Innovation and Productivity
Thank You!

Content and Causality in Influence Networks

Sinan Aral
NYU
@sinanaral
Viral state is correlated with application activity. However, this relationship disappears when you control for number of adopters in your local network. Number of adopters is still correlated with activity even when you control for use of viral features.
Viral Features, Adoption and Use

Feature
Existence

Peer Adoption

Use

Omitted Variables

Variables
Adjudicating Alternate Explanations

- Correlation between Peer Adoption & Application Use could be evidence of Network Effects.
- It could also be explained by:
  - Unobserved Heterogeneity / Omitted Variables
  - Demand Effects – Existence of Features
- But, treatment is randomized so omitted variables cannot explain the variation we see in adoption and use.
- There could be an interaction effect between an omitted variable and a feature itself, but we observe a correlation between peer adoption and application use controlling for feature use.

1. (a) and (b) inconsistent with the discrepancy in app use between treatment groups.
2. If Demand Effects: Controlling for them should remove any spurious correlation between peer adoption and use.
3. Controlling for feature use, peer adoption still highly correlated with app use.
Date: Sun, 01 Feb 2009 10:02:10 -0500
From: xxx@yyy
To: abc@123
CC: zzz@yyy, abc@zzz
Subject: Re: IWP 1 Extended Thoughts

Actually,

...be even more succinct about 3 main take aways:

1. Given our ability to make connections between abstract concepts, our productivity is determined more by our ability to multitask, than by our ability to conduct sequential work faster.

So, let's explore the mechanisms behind multitasking a bit more:

2. The relationship between output and multitasking is convex at low levels of multitasking and concave at high levels of multitasking. Because information inputs are non-rival and complementary, unlike physical inputs, their use enables convexity in the relationship between multitasking and output at low levels of multitasking. Because human information processing is constrained by bounded rationality, and limited cognitive capacity the relationship between multitasking and output is concave at higher levels of multitasking.

So, how do we acquire the inputs we use? Socially:

3. Efficient positioning in the social network creates efficient means to gather and use information and is correlated with higher productivity. Because we require a social support system of information acquisition (embodied in our social networks) which we rely on to extend our own individual mental capacity, we gather information inputs socially (and through IT which we use as a control variable).

Dr. XXX

Construction of E-mail Vectors

1. Header information is extracted to create the social network. Names are matched and identities are validated by hand.
2. The subject and body of the e-mail message are analyzed to extract frequencies of use of keywords (Steps 3-6).
3. Stop words (e.g. “a,” “an,” “the,” “and,” and other common words) with high frequency across all e-mails are removed as shown by words that have been struck.
4. Keywords are extracted based on the three principles outlined on pages 30-31 of the manuscript.
5. Keywords are root-stemmed, such that for example “multitask,” “multitasking,” become “multitask*.”
6. The frequency of each key word is counted and recorded.
7. A vector representing the e-mail is created which logs: the e-mail ID, the ID number of each keyword used and the frequency of use of each keyword noted inside brackets as follows:
   
   E-mailID(7842B/748821<9>; ...; 849247<2>)
   
   A vector representing the example e-mail to the left is shown in truncated form below.
8. The content similarity of e-mail vectors is then compared using several standard distance metrics such as the Cosine distance.

\[
\vec{d}_j = (IWP < 1>; connection* < 1>; productivity < 2>; multitask* < 9>; sequential < 1>; output < 3>; ...; input* < 4>; social* < 5>; IT < 1>; control < 1>; variable < 1>)
\]
The diversity of the information in an email inbox or outbox is measured by calculating the Cosine Distance of each email vector $d_i$ to the mean vector of that inbox or outbox $M_{in}$, and then averaging across all emails:

$$CosDist = 1 - \cos(d_i, M_i)$$

$$ID_i = \frac{\sum_{j=1}^{N} (1 - \cos(d_{ij}, M_i))^2}{N}$$


“Creating Social Contagion through Viral Product Design: A Randomized Trial of Peer Influence in Networks.” Forthcoming in *Management Science*

“The Diversity-Bandwidth Tradeoff” Forthcoming in *American Journal of Sociology*
Five Broad Research Questions

1. What is peer influence (formally)?

2. How do characteristics of the product or behavior affect peer influence and contagion?
   – What characteristics make a viral video go viral?
   – How do we design viral products?
   – How do network externalities inherent in a product affect its diffusion?

3. What is the role of sustained use in creating sustainable contagions?
   – Do users need to adopt and maintain interest or use or just adopt once?
   – What is the role of churn? Engagement?

4. How do distributions of individual characteristics over network nodes affect contagion?
   – How does assortativity / homophily affect propagation of influence?
   – Do persuasive individuals tend to be of high degree?
   – Do influential tend to be surrounded by susceptibles or do influential cluster?

5. Are there ‘systems’ of complementary contagion management strategies?
   – How do referral incentives and targeting interact? Are they complements or substitutes?
We specify three types of treatments that we refer to as the Viral Feature State of a user:

- **Non-Viral User** ($Vi = 0$): Both passive and active viral features are disabled (baseline)
  
  [5% of total user population]

- **Passive Viral User** ($Vi = 1$): Only passive viral features (e.g., notifications) are enabled

  [47.5% of total user population]

- **Active Viral User** ($Vi = 2$): Both passive viral features and active viral features (e.g., invites) are enabled

  [47.5% of total user population]
Findings

1. **Viral Product Design Features** produce econometrically identifiable peer influence and social contagion effects.

2. **Active-Personalized Viral Features** are more effective per message but are used much less often and so produce less total peer adoption in the network than **Passive-Broadcast features**.

3. **Our data are consistent with the existence of positive network externalities** that drive a feedback loop of peer adoption and sustained product use.

4. **Viral Feature** outperform traditional banner ads and email campaigns in generating adoption.
## Table 4: Variance-Corrected Proportional Hazards of Contagion in Networks of Baseline, Passive and Active Treatment Groups

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Hazard Ratio (SE)</strong></td>
<td><strong>Hazard Ratio (SE)</strong></td>
<td><strong>Hazard Ratio (SE)</strong></td>
<td><strong>Hazard Ratio (SE)</strong></td>
</tr>
<tr>
<td>Viral State =</td>
<td><strong>3.46</strong>* (1.18)</td>
<td><strong>3.35</strong>* (1.15)</td>
<td><strong>2.50</strong> (.86)</td>
<td><strong>2.51</strong> (.86)</td>
</tr>
<tr>
<td>Passive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Viral State =</td>
<td><strong>4.44</strong>* (1.64)</td>
<td><strong>4.21</strong>* (1.56)</td>
<td><strong>3.33</strong>* (1.24)</td>
<td><strong>3.31</strong>* (1.24)</td>
</tr>
<tr>
<td>Active</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Application</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Notifications</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Invites</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-4694.359</td>
<td>-4631.795</td>
<td>-4544.845</td>
<td>-4542.577</td>
</tr>
<tr>
<td>X² (d.f)</td>
<td><strong>19.34</strong>* (2)</td>
<td><strong>57.41</strong>* (3)</td>
<td><strong>298.78</strong>* (4)</td>
<td><strong>307.47</strong>* (5)</td>
</tr>
<tr>
<td>Observations</td>
<td>3929</td>
<td>3929</td>
<td>3929</td>
<td>3929</td>
</tr>
</tbody>
</table>

Notes: ***p<.001; **p<.05; *p<.10;
<table>
<thead>
<tr>
<th>Table 5: Click Stream Analysis of Responses to Viral Messages and Adoption</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td><strong>Messages Sent</strong></td>
</tr>
<tr>
<td>Invitations</td>
</tr>
<tr>
<td>Notifications</td>
</tr>
</tbody>
</table>
Invites are a form of **Viral Messaging**

In addition to viral messaging, Facebook applications also make use of **traditional online advertisements** by placing ads directly inside the application region.

There is a market for Within-Application advertising.

automatically when a user takes an action within an application. They are delivered to a user’s Facebook friends like this.
<table>
<thead>
<tr>
<th>Feature</th>
<th>Personal Invitations</th>
<th>Passive Awareness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influence Per Message</td>
<td>↑ 6%</td>
<td>↑ 2%</td>
</tr>
<tr>
<td>Global Diffusion</td>
<td>↑ 98%</td>
<td>↑ 246%</td>
</tr>
<tr>
<td>Stickiness</td>
<td>↑ 17%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>----------------</td>
<td>----</td>
<td>-----</td>
</tr>
<tr>
<td></td>
<td>Beta</td>
<td>Beta</td>
</tr>
<tr>
<td></td>
<td>(SE)</td>
<td>(SE)</td>
</tr>
<tr>
<td>Viral State =</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passive</td>
<td>.129*</td>
<td>.112</td>
</tr>
<tr>
<td></td>
<td>(.074)</td>
<td>(.079)</td>
</tr>
<tr>
<td>Viral State =</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active</td>
<td>.190***</td>
<td>.171**</td>
</tr>
<tr>
<td></td>
<td>(.074)</td>
<td>(.079)</td>
</tr>
<tr>
<td>Degree</td>
<td>-.0001</td>
<td>-.0001</td>
</tr>
<tr>
<td></td>
<td>(.0001)</td>
<td>(.0001)</td>
</tr>
<tr>
<td>Facebook</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity</td>
<td>.054***</td>
<td>.042***</td>
</tr>
<tr>
<td></td>
<td>(.016)</td>
<td>(.015)</td>
</tr>
<tr>
<td>Application</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Notifications</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Invites</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adopters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F Value</td>
<td>3.51***</td>
<td>4.87***</td>
</tr>
<tr>
<td>(d.f)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.002</td>
<td>.003</td>
</tr>
<tr>
<td>Observations</td>
<td>6310</td>
<td>5766</td>
</tr>
</tbody>
</table>
| Notes: ***p<.001***
**Adjudicating Alternate Explanations**

- **Correlation between Peer Adoption & Application Use could be evidence of Network Effects.**
- **It could also be explained by:**
  - *Unobserved Heterogeneity / Omitted Variables*
  - *Demand Effects – Existence of Features make Product more interesting to use.*
- **But, treatment is randomized so omitted variables cannot explain the variation we see in adoption and use across treatment groups.**
- **There could be an interaction effect between an omitted variable and a feature itself, but we observe a correlation between peer adoption and application use controlling for feature use.**

(a) (a) and (b) inconsistent with the discrepancy in app use between treatment groups.
(b) If Demand Effects: Controlling for them should remove any spurious correlation between peer adoption and use.
(c) Controlling for feature use, peer adoption still highly correlated with app use.
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean (SD)</strong></td>
<td><strong>Min</strong></td>
<td><strong>Max</strong></td>
<td><strong>N</strong></td>
<td></td>
</tr>
<tr>
<td>.0002 (.0001)</td>
<td>.0001</td>
<td>.001</td>
<td>523</td>
<td></td>
</tr>
<tr>
<td>.002 (.001)</td>
<td>.001</td>
<td>.013</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>.015 (.024)</td>
<td>.005</td>
<td>.14</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>.034 (.010)</td>
<td>.021</td>
<td>.054</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>.046 (.008)</td>
<td>.037</td>
<td>.067</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>.099 (.044)</td>
<td>.053</td>
<td>.14</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
The purpose of WIN: “to bring together leading researchers studying ‘information in networks’ to build a lasting multidisciplinary research community.”
Influence Prompted by Viral Messaging
@sinanaral

Application 1

Welcome, Sinan

INVITE FRIENDS

Facebook Home Page

Install Application?
Welcome, Sinan
INVITE FRIENDS

Welcome, Lev
<table>
<thead>
<tr>
<th></th>
<th>Baseline (N=405)</th>
<th>Passive (N=4600)</th>
<th>Active (N=4682)</th>
<th>t-statistic (B-P)</th>
<th>t-statistic (B-A)</th>
<th>t-statistic (P-A)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>t-statistic (SE)</td>
<td>t-statistic (SE)</td>
<td>t-statistic (SE)</td>
</tr>
<tr>
<td>Age</td>
<td>31.51 (13.80)</td>
<td>30.81 (13.31)</td>
<td>29.94 (13.27)</td>
<td>.46 (13.35)</td>
<td>1.03 (13.31)</td>
<td>1.45 (13.24)</td>
</tr>
<tr>
<td>Gender (1=Male)</td>
<td>.25 (.44)</td>
<td>.33 (.47)</td>
<td>.32 (.47)</td>
<td>-1.57 (.47)</td>
<td>-1.42 (.46)</td>
<td>.40 (.47)</td>
</tr>
<tr>
<td>Degree†</td>
<td>171.79 (223.88)</td>
<td>170.25 (278.64)</td>
<td>166.97 (248.77)</td>
<td>.09 (275.13)</td>
<td>32 (247.15)</td>
<td>.55 (263.82)</td>
</tr>
<tr>
<td>Number of Facebook Wall Posts</td>
<td>40.52 (79.89)</td>
<td>36.45 (94.16)</td>
<td>37.07 (246.76)</td>
<td>.46 (93.11)</td>
<td>.15 (238.20)</td>
<td>-.09 (188.31)</td>
</tr>
<tr>
<td>Table 3. Summary Statistics and Mean Comparisons of Active, Passive and Baseline Users</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td><strong>Baseline (N = 405)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Passive (N = 4600)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Active (N = 4682)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mean (SD)</strong></td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>t-statistic (SE)</td>
<td>t-statistic (SE)</td>
<td>t-statistic (SE)</td>
</tr>
<tr>
<td><strong>Number of Adopters in User’s Local Network</strong></td>
<td>.01 (.12)</td>
<td>0.07 (.35)</td>
<td>0.10 (.44)</td>
<td>-2.84*** (.34)</td>
<td>-3.60*** (.43)</td>
<td>-3.64*** (.40)</td>
</tr>
<tr>
<td><strong>Percentage of Adopters in User’s Local Network</strong></td>
<td>.02 (.002)</td>
<td>.09 (.01)</td>
<td>.15 (.01)</td>
<td>-1.92* (.01)</td>
<td>-2.35** (.01)</td>
<td>-2.83*** (.01)</td>
</tr>
<tr>
<td><strong>Maximum Diffusion Depth</strong></td>
<td>.01 (.11)</td>
<td>.04 (.22)</td>
<td>.05 (.24)</td>
<td>-2.53* (.21)</td>
<td>-3.01*** (.24)</td>
<td>-1.98*** (.23)</td>
</tr>
</tbody>
</table>
Creating Social Contagion through Viral Product Design

A Randomized Field Experiment

Sinan Aral, NYU Stern & MIT  Dylan Walker, NYU Stern
Randomized Peer Influence Experiments in Online Social Networks

WISE 2009
Phoenix, AZ

Sinan Aral, NYU Stern & MIT
Dylan Walker, NYU Stern
Randomization of Passive Viral Messages

- Notifications are typically one-to-many messages.
- The recipients of notification are randomly selected from the set of the sender’s social network peers.

A new set of random recipients are selected for each notification.

- We record when notifications are sent and to whom.
- We record recipient responses to notifications:
  - If/when the user clicks on the notification, leading to an application installation page.
  - If/when the user proceeds to install the application.
Hypothesis 1 (H1). *Enabling viral product design features increases the likelihood of adoption among peers of current users.*

Hypothesis 2 (H2). *Viral product design features that require more activity on the part of the user and are more personalized to recipients create greater marginal increases in the likelihood of adoption per message.*

Hypothesis 3 (H3). *Viral product design features that require more activity on the part of the user and are more personalized to recipients generate fewer total viral messages.*
Influence conditional on Viral Features
- Active Viral Messages
- Passive Viral Messages
- Traditional Advertising

Influence conditional on Viral Features and
- Characteristics of Sender
- Characteristics of Recipient
- Characteristics of Dyad
- Network Properties of Sender
- Network Properties of Recipient
- Network Properties of Dyad
Influence conditional on Sender Attributes and Viral Features

\[ F(\text{friends of } i \mid V_i, \{X_i\}) \]

Fraction of friends that adopt Application 2 conditional on the Viral Feature State of user \( i \) and the attributes of user \( i \)

Notice, this measure will pick up if some attributes of user \( i \) make him more or less influential
Susceptibility to Influence via Viral Messaging

\[ \{X_j\} \quad \text{Attributes of friend } j \]

\[ R_j = 1 \quad \text{iff friend } j \text{ received a viral message} \]

\[ F(\text{friends } j \mid R_j = 1, \{X_j\}) \]

Fraction of all friends of some Application 1 user that adopt Application 1 conditional on the receipt of a viral message and the attributes of friend

Notice, this measure will pick up if some attributes of a friend \( j \) make him more or less easily influenced
Experiment 2: Randomized Traditional Advertising

Randomly showed traditional banner ads to some (experimental) application users
Experiment 2: Randomized Traditional Ads

Control Group

Experimental Group

Experimental Group
Sees a Banner Ad for Application 2

Control Group
Sees No Banner Ad
Experiment 2: Randomized Traditional Ads

We then compare:
- click throughs
- adoption
- usage data

of neighbors of:
- Experimental Group
- Control Group

Allows us to test:
1. Average Treatment Effect of Traditional Advertising on Peer Adoption and Network Propagation
2. A focus on latent communication channels
II. Influence Prompted by Traditional Advertising

@sinanaral

Hey Dylan, have you heard of this “Application 2”? It gets you citations.

I like citations. I’ll try it out.

As in this example, some channels of communication between friends are latent.
Facebook Applications

Facebook applications

- are independently developed web applications
- run within the framework of Facebook
- are integrated with the social functions of Facebook:
  - can utilize the inter-user communication channels and knowledge of the social network provided by Facebook,

- When a user elects to adopt or install a Facebook application
  - grants the application access to:
    - the user’s immediate social network
    - the user’s profile data and the profile data of their immediate social network peers (according to their privacy settings)
Notice, that when Prof. Mendelson visits Application 1, it doesn’t show him an advertisement, because he’s not in the treatment group

This idea can be generalized.

You can imagine using Application 1 as a delivery system for ultra-targeted advertisements, or to select treatment populations based on specific user attributes.

I won’t discuss that here.

We use Application 1 to deliver advertisements for a second application to randomly selected users.

We then observe the Application 2 adoption amongst the peers of Application 1 users.
We perform both experiment I and II simultaneously, because application users are a precious resource.

The possible treatment an Application 1 user receives are given by:

<table>
<thead>
<tr>
<th></th>
<th>Non Viral</th>
<th>Passive Viral</th>
<th>Active Viral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad</td>
<td>2.5%</td>
<td>23.75%</td>
<td>23.75%</td>
</tr>
<tr>
<td>No Ad</td>
<td>2.5%</td>
<td>23.75%</td>
<td>23.75%</td>
</tr>
</tbody>
</table>
Potential Issues: Viral Adoption Chains

User has installed Application 1

User is in Experiment Group

User is in Control Group

User sends viral message to Friend

Friend installs

Friend → User

Are there differences between:

- Users who became users through “initial advertising”
- Users who became users through viral messaging
At least two levels of potential selection bias:

1. Selection into experiment – Adoption of App 1.
   - Here we are careful about how representative our sample is of the population of Facebook users.
   - We can test observable characteristics.
   - But there may be unobserved differences.

2. Selection into the treatment – sending of viral messages (choosing to actively invite friends)
   - Here we are careful about whether those who choose to message are systematically different.
   - We test the average treatment effect of having messaging turned on – that’s random.
   - This is analogous to traditional selection bias in randomized trials. There are known tests and methods
Whether users choose to invite and whether friends of users respond depend on the applications we use.

- More engaging, more naturally “viral” applications may induce different usage behavior and different responses from recipients of viral messages.
- Some applications have network externalities.
- We use representative, engaging, real world applications (not developed solely for the experiment).
- We carefully consider how our applications compare to others e.g. in how viral they are, if they have network externalities etc.
Potential Issues

Friends of multiple treated users

We group peers with n treated friends together for modeling.

Peers are not double counted

Friends of users in experiment and control

Depending on the nature of the treatment, simultaneous peers of users in the experiment and control group may be unclearly classified.

In some treatment cases, simultaneous relationships control group users is of no consequence.
• Which viral message are users responding to?

• It could be the case that some peers respond positively to a viral message only after they have received multiple viral messages from the same or many different friends.

• Even when a user doesn’t click on a viral message, they may be affected by it.
  – Complex Contagion Model (Centola and Macy 2005)

  – Because we observe when viral messages are received, we can distinguish 1st, 2nd, … , Nth time recipients receive a viral message (and from whom) and classify them accordingly.
Causal Statistical Estimation

I used to think correlation implied causation.

Then I took a statistics class. Now I don’t.

Sounds like the class helped.

Well, maybe.
Experiments are Running As I Speak

Stay Tuned!

Thank You!!
Questions?
We have to be very aware of exactly what the response is conditional on and characterize results accordingly…

Examples:

- Profile Box – Influence conditional on installing the box and ‘recipients’ surfing the home page of the user.
Potential Issues – Exact Conditional Probabilities

- We have to be very aware of exactly what response is conditional on …
- Examples:
  - Profile Box – Influence conditional on installing the box and ‘recipients’ surfing the home page of the user.
Potential Issues – Selection Effect on Peer Response

- We have to be very aware of exactly what the response is conditional on and characterize results accordingly…

- Examples:
  - Profile Box – Influence conditional on installing the box and ‘recipients’ surfing the home page of the user.
Current Status of Experiments
@sinanaral

• We have data on millions of Facebook users
• experiment has been running for 1 week
• we have 3500 application users, hundreds of thousands of peers
• Thousands of viral messages have been sent
• we plan to continue the experiment for 5-6 weeks
• stay tuned
Acknowledgements

• Collecting data of this scale involves significant technical challenges
  – Hundreds of Gigabytes of data being collected live
  – Weeks of data processing
  – This would not have been possible without the “Stern Cloud” created and operated by the Stern Center for Research Computing
  – We are indebted to Norman White for his extremely timely assistance
Privacy and Human Subjects Safeguards

1) Privacy safeguards
   – Facebook users are given a large number of privacy options that control the visibility of their data
   – Facebook users explicitly grant permission to applications
   – All user data that we collect are de-identified

2) Experimental data on human subjects
   – Our experimental application features involve minimal harm
   – Our experiments are analogous to A/B tests involving application features. Standard tests such as these are performed regularly by online service vendors to improve user experience.

3) Facebook Terms of Service
   – In accordance with Facebook TOS, direct profile data is not retained
   – We retain and work with derived data only
Extra Slide: Profile Data

- **Demographics**
  - age, sex, current and hometown location, education and employment history,

- **Relationships and Social Interactions**
  - Facebook friends, family members, relationship status to significant other, social group participation, photo co-appearance

- **Preferences**
  - religious views, political views, activities, interests,

- **Product Tastes**
  - movies, music, book, tv shows, Facebook application adoption
To study the influence that a Facebook user exerts on his peers, we randomly split all users of Application 1 into two groups:

- Users in the **Experimental Group** are subjected to the treatment while users in the **Control Group** are not.

We then examine the behavior of the friends of users in both groups.
In this example it appears that females are relatively unaffected by viral messages, whereas males exhibit a greater likelihood to adopt when they receive a viral message.
Randomized Trials of Peer Influence

**Experiment 1:** Randomized Viral Messaging

Randomly Enabled Some (Experimental) Application Users to send Viral Messages to their Friends

**Experiment 2:** Randomized Traditional Advertizing

Randomly Showed Traditional Banner Ads to Some (Experimental) Application Users

Compared *Adoption and Usage Behavior of the Neighbors of Experimental Groups 1 and 2 to the Neighbors of:*

*Control Groups:* No Messaging; No Ads
Two Experiments

I. Influence arising from exchange of viral messages
   - Treated users are given the ability to send viral messages within Application 1
   - These viral messages encourage the recipient to install Application 1
   - Observe peer adoption of Application 1

II. Influence arising from exposure to traditional advertisement
   - Treated users receive an advertisement for Application 2 that is displayed within Application 1
   - Observe peer adoption of Application 2
Users can add a box to their profile with information pertaining to this application. It looks like this:

Friends that browse a user’s profile and click on this box are taken to a page where they are given the opportunity to install the application.
Users can invite friends to install this application and include a personal message.
Viral Communications

• We classify viral communications into two categories

  Passive viral messages
  Impersonal messages that are generated automatically by an application as a consequence of an application user’s activity
  e.g., profile box, notifications

  Active viral messages:
  Direct solicitations by a user to his friends to adopt an application, join a group or engage in an application-specific behavior
  e.g., invites
“Causal Knowledge Requires Controlled Variation.”
– Armin Falk and James Heckman, Science (2009)

Comparing treated to untreated without random assignment creates bias in the potential untreated outcomes of treatment and comparison groups.

Randomization: treatment and control groups differ in expectation only through their exposure to the treatment.
Viral state is correlated with application activity. However, this relationship disappears when you control for number of adopters in your local network.

Number of adopters is still correlated with activity even when you control for use of viral features.
Viral Features, Adoption and Use

- Feature Existence
- Peer Adoption
- Use

Omitted Variables
Correlation between Peer Adoption & Application Use could be evidence of Network Effects.

It could also be explained by:
- Unobserved Heterogeneity / Omitted Variables
- Demand Effects – Existence of Features

But, treatment is randomized so omitted variables cannot explain the variation we see in adoption and use.

There could be an interaction effect between an omitted variable and a feature itself, but we observe a correlation between peer adoption and application use controlling for feature use.

Adjudicating Alternate Explanations

1. (a) and (b) inconsistent with the discrepancy in app use between treatment groups.
2. If Demand Effects: Controlling for them should remove any spurious correlation between peer adoption and use.
3. Controlling for feature use, peer adoption still highly correlated with app use.
Baseline Hazards Increasing in $k$ adoption events

$y = 0.0002e^{3.4939k}$
$R^2 = 0.9875$

$\lambda_{0k}$

$y = 0.0002e^{1.1785k}$
$R^2 = 0.8956$