Social Ties and User Generated Content: Evidence from an Online Social Network

Reto Hofstetter
University of Bern

Scott K. Shriver
Stanford GSB

Harikesh S. Nair
Stanford GSB

* The Networks, Electronic Commerce, and Telecommunications (“NET”) Institute, http://www.NETinst.org, is a non-profit institution devoted to research on network industries, electronic commerce, telecommunications, the Internet, “virtual networks” comprised of computers that share the same technical standard or operating system, and on network issues in general.
Social Ties and User Generated Content: Evidence from an Online Social Network

Reto Hofstetter*  Scott K. Shriver†  Harikesh S. Nair‡

First version: November 2009
This version: April 13, 2010§

Abstract

We use variation in wind speeds at surfing locations in Switzerland as exogenous shifters of users’ propensity to post content about their surfing activity onto an online social network. We exploit this variation to test whether users’ social ties on the network have a causal effect on their content generation, and whether content generation in turn has a causal effect on the users’ ability to form social ties. Economically significant causal effects of this kind can produce positive feedback that generate multiplier effects to interventions that subsidize tie formation. We argue these interventions can therefore be the basis of a strategy by the firm to indirectly facilitate content generation on the site. The exogenous variation provided by wind speeds enable us to measure this feedback empirically and to assess the return on investment from such policies. We use a detailed dataset from an online social network that comprises the complete history of social tie formation and content generation on the site. The richness of the data enable us to control for several spurious confounds that have typically plagued empirical analysis of social interactions. Our results show evidence of significant positive feedback in user content generation. We discuss the implications of the estimates for the management of the content and the growth of the network.

*Assistant Professor of Marketing, University of Bern, Email: Reto.Hofstetter@imu.unibe.ch
†Doctoral Candidate in Marketing, Stanford GSB, Email: scott.shriver@gsb.stanford.edu
‡Associate Professor of Marketing, Stanford GSB, Email: harikesh.nair@stanford.edu.
§We thank the NET Institute (www.NETinst.org) for supporting this research. All errors are our own.
1 Introduction

Social networking sites, such as Facebook or MySpace, provide platforms for users to communicate and to connect with one another. Social networking sites are increasingly relevant in the online economy. Facebook, for example, has grown from about 18 million unique visitors in September 2007 to 39 million in September 2008, which represents a 116% increase (Nielsen 2008a). The market research firm IDC reports that more than half of U.S. consumers with Internet access use social networking sites. Despite the rising importance of online social networks, academic studies of user behavior on these sites has been limited (e.g. Trusov et. al. 2009).

A key challenge facing social networking sites is monetization strategy. The dominant monetization model for online social networks is advertising. In 2008, U.S. social network advertising expenditures reached $1.2 billion and are expected to grow to $1.6 billion by 2013 (eMarketer 2008b). Advertising in a social network is linked to the volume of page views on the network, which is a function of users’ social activity on the site. The social activity is ultimately linked to consumption and creation of online content. While such user-generated content is critical to the site’s revenue model, the management of the content is becoming increasingly complex as firms have relatively few practical levers at their disposal to induce users to post content. One source of the difficulty is that users’ content generation habits largely reflect personal activities and tastes, and may not be significantly influenced by features the site can control, such as provision of better tools for blogging or uploading photos. Additionally, much of the content is typically posted by a small proportion of the users (for example, in our data 10% of the users account for 80% of the generated content). One option for the website is to selectively provide access to sophisticated tools for these groups of users. However, competition amongst social networks has typically precluded discriminating amongst users on the basis of access to tools. Rather, the trend in the industry is to provide comprehensive access to tools to the entire universe of users, so as to attract the largest installed base of users. Finally, subsidizing users to generate content is also not very feasible. For instance, directly paying users to generate content can bias the character of the content, which may have negative repercussions in many situations1. In essence, one conundrum for social media sites is that when the site operator tries to influence the user generated space, content contribution loses its appeal for users who want to be in control.

We consider a separate mechanism by which the firm may indirectly manage the genera-

---

1For example, WalMart stirred controversy by paying Edelman PR to create positive blog listings for its products (Glaser 2006). For a formal analysis, see Friestad and Wright 1994 and Verlegh et al. 2004.
tion of content on its site. Our main insight is driven by the fact that the motivation of users to post content is derived from the social value of its consumption: a user posts content so he obtains the social benefit of its consumption by his friends. Analogously, he spends time on the website so as to obtain the social value of consuming the content generated by others. Both imply that content generation is a function of the extent of social ties the user enjoys on the website. Thus, an indirect way for the firm to promote content generation is to facilitate social tie formation, thereby influencing the users’ network structure. If the network structure has a true causal effect on content generation, then the firm benefits from policies that subsidize link formation. Several policies that encourage friendship formation are already available in extant social networks. These include providing online privileges based on the number of friends a user has or providing information provision to users about the tastes, profiles and activities of potential friends. In this paper, we demonstrate that such efforts are also an effective means to indirectly facilitate content generation.

The goals of this paper are threefold. First, we test whether network link formation has a causal effect on the production of user content on an online social network. Testing this proposition is confounded by the fact that unobservables that drive content generation may also drive link formation, thereby generating spurious correlation. The main identification challenges are endogenous group formation, correlated unobservables and simultaneity (e.g. see Manski 1993; Moffit 2001). We discuss how we overcome these challenges in our analysis. Second, we ask whether an analogous causal effect exists for the reverse process, whereby tie formation is driven by content generation. This is plausible as users who generate more content may be more frequently invited to form social ties with others. We are interested in assessing the reverse effect because of the potential for social multipliers which can significantly boost the effect of marketing interventions (e.g. Nair et. al. 2006). In particular, if both causal effects are strong, the interaction between content production and network linkage generates a feedback that accentuates the benefits to content generation from subsidizing tie formation. Only causal effects can result in such social feedback. Hence, uncovering causal effects accurately is key to formulating policy. This in turn relies on a credible sorting of other sources of correlation in observed behavior from true causal phenomena. Our third goal is to derive implications of both causal effects for the management of content on the website. In particular, we seek to measure the influence of such feedback effects on the return on investment (ROI) from marketing efforts to increase social tie formation.

We address these research questions using rich, detailed data from an online social network based in Switzerland, named Sourlrider.com. Sourlrider is one of the largest sports-
based communities in Switzerland, and is primarily focused on windsurfers. Users post content on Soulrider about their surfing activities, including blogs that report wind speeds and conditions at specific surfing locations. Our data comprises the complete history of the connectivity between users, user demographics as well as user content creation on Soulrider. Of particular importance to our identification strategy, we also have access to the time series of this information for the universe of users of the network. The panel aspects of the data enable us to use within-member variation in the joint distribution of network linkage and content generation in order to identify the causal effects while controlling for endogenous group formation via user fixed effects. Additionally, the panel data enable the incorporation of time fixed effects to control for common unobservables that drive content generation and friend formation similarly. Further, we augment the website data with high frequency information on wind forecasts at all surf locations in the country. Surfing is possible for most users only if wind-speeds are greater than or equal to 4 BFT. We use the variation in wind speeds as exogenous shifters of users’ propensity to visit surf locations, and to subsequently post content about their surfing activity. We show that wind speeds significantly explain the observed variation in content postings on the website. The wind speeds serve as instruments for content generation, thereby enabling us to identify both feedback effects above. We present several tests of the validity of the identification conditions.

Our analysis provides evidence for significant causal effects in both directions. We find that the number of social ties has a positive effect on a user’s content generation, and that content generation in turn has a significant positive effect on forming social ties. Our results are robust across a variety of estimators that transparently incorporate the sources of identification discussed above. We find large effects of controlling for endogenous group formation and correlated unobservables. In the absence of these controls, we find measures of social effects are overstated about 25-30% on average across model specifications. We use our estimates to measure the revenue implications of augmenting social ties on the network. Specifically, we measure how much incremental advertising revenue the site may earn by facilitating a unit increase in the number of social ties amongst its users. Our results suggest that adding social ties has large effects on content generation, and consequently on page impressions and ad-revenue. Our results suggest that incorporating multiplier effects is critical to obtaining an accurate picture of the ROI from such activity. In particular, we find that in the absence of accounting for such multipliers, the ROI from subsidizing link formation is estimated to be 10.2% lower on average, and as much as 40% for some users.

\footnote{BFT stands for “Beaufort,” the international wind scale used in weather reporting.}
The large differences accentuate the need for the correct identification of subsets of users who respond to targeting, and the correct measurement of the ROI from interventions that induce such users to post more content.

The remainder of the paper is organized as follows. In section 2, we provide a short overview of the online social network Soulrider.com, from which the data are collected. We also describe our data in detail. In section 3, we present a variety of estimators for causal effects. Section 4 presents the results, and Section 5 discusses the implications of our model for website ROI. Section 6 concludes.

2 Data and Model-free Evidence

2.1 Soulrider.com

Soulrider.com is a privately held website that focuses on extreme sports such as wind-surfing, surfing and snowboarding. It was founded in 2002, and is based in Europe. As of December 2008, Soulrider.com had a total of 6,217 registered users. Registered users self-report that they learn about Soulrider.com through word-of-mouth (WOM) (37.35%), search engines (29.62%), external links (16.66%), E-mail invitation (5.66%), or other reasons (10.70%).

Users of Soulrider.com can consume both existing content and submit their own. Most content on the website, such as blogs or forum messages, are generated by users themselves. Other content, such as national level sport industry news, is provided by third party contributors and is not affected by users. Users who wish to post content or engage in social networking activities are required to create a free account on the website. The main value of content is informational. Blogs and posts typically contain information about where surf conditions are best; where other users plan to surf in a given week; and reports on wind-speeds at those locations. These help users better organize and coordinate their sporting activities.

In addition to consuming and generating content, users can create ties to other users and thus, take part in an online social-network. Social networking of the users is facilitated by the website through various functions such as a people search-engine, an E-mail invitation tool, interest groups, and an “add as friend” function that handles the mechanics of creating ties in the online interface. Communication among users is eased by internal mail functionality, instant messaging, and public chat.

The website is representative of networks targeting young adults. In of end of 2008, 77% of the users on the website were male and 23% female. The mean user is 30.2 years old, logged on to the website 5.5 times per month, generated 64.6 page impressions (PI-s) per month, and spent about 140 seconds on the website per visit. The social network
grew by 1,687 (37.2%) users in 2008. Of all users, 1,592 (25.6%) added at least one friend, which generated 2,305 new social ties. The mean user possesses 1.5 friends. Content was contributed by 2,161 (34.7%) users.

The website counts an average of about 50,000 visits per month. Visitors stay on the website for 4 minutes on average and generate approximately 400,000 page impressions per month. Google has indexed 39,900 pages and 161 backlinks for Soulrider.com and page-ranks it with a value of 5 (as of December 20th, 2008). Thus, to summarize, Soulrider.com is a medium-sized social network, appealing to a core community that has shared interests and strong incentives to maintain ties, and which grows primarily by word of mouth.

2.2 Data

We worked with Soulrider.com to add a logging functionality to capture all possible activities that might occur on the website concerning the development of the social network and the generation and consumption of content. Our data comprise complete details of users’ ties and content generation from May 3rd, 2009 to October 4th, 2009, a period spanning 32 weeks. For the purposes of this paper, we focus on the group of 368 self-identified wind-surfers on the website. Thus, our panel is of size 11,776 observations (368 users × 32 weeks). We operationalize user content via a generic variable we call blogs, which counts the number of postings the user has made to the website each period. A posting is counted as adding 1 to the blog variable if it contains any text or photos. Thus, if a user posted a photo and a text message on the website, two text messages, or two photos, the blogs variable would equal 2. To test causal effects in this paper, we do not account for the volume or type of content added, but focus on the incidence of postings. We operationalize ties by a variable, friends, which counts the new friendship links requested by others to each user per period. Essentially, we ask if content creation facilitates new friend requests, and whether new friend requests in turn facilitate the creation of new content.

2.2.1 Basic patterns in the data

We now discuss stylized patterns in the data to motivate our subsequent model development. We structure this discussion as follows. First, we describe key patterns in the generation of content and the pattern of social ties. Second, we check for interrelationships in content generation and network structure, which is linked to the key goals of this research. Further, we present evidence supporting the identification provided by exogenous wind forecast data on content generation.

---

3Impression statistics from Google Analytics for the 30 day period ending December 20th, 2008.
Table 1: Descriptive Statistics of Blogs and Friends

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>blogs</td>
<td>11776</td>
<td>0.12</td>
<td>0.457</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>friends</td>
<td>11776</td>
<td>0.05</td>
<td>0.239</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 2: Joint distribution of friends and blogs

<table>
<thead>
<tr>
<th>friends</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>blogs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>10,412</td>
<td>346</td>
<td>28</td>
<td>2</td>
<td>1</td>
<td>10,789</td>
</tr>
<tr>
<td>1</td>
<td>617</td>
<td>57</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>684</td>
</tr>
<tr>
<td>2</td>
<td>182</td>
<td>27</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>217</td>
</tr>
<tr>
<td>3</td>
<td>45</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>52</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>11,284</td>
<td>438</td>
<td>47</td>
<td>6</td>
<td>1</td>
<td>11,776</td>
</tr>
</tbody>
</table>

Friends and Blogs

We start by presenting the distribution of friends and blogs for the set of users in the data. Table (1) presents the descriptive statistics for the blogs and friends variables. On average, users post about 0.1 blogs per week (Max 6), and add about 0.05 friends per week (maximum 4). There are also a large number of user-weeks when no blogs are posted, or no friends are added.

The joint distribution of friends and blogs is summarized by the cross-tabulation in Table (2). There is a large mass point at zero. The across-consumer, across-time correlation between the variables is positive and significant ($\rho = 0.13$). The fact that outcomes are discrete and limited in range suggests that a count model is appropriate for modeling these data.

Social Ties and exogenous variation in wind

The challenge in the empirical analysis is to identify the causal effects of blogs on friends, and of friends on blogs from these data. Our strategy for obtaining these causal effects relates to the differences in the position of users of Soulrider.com across Switzerland, and the resulting proximity of these users to different surfing locations. In particular, users that are close to viable surfing locations are likely to visit those locations often, prompting them to blog conditions at these locations frequently. Hence, across geographic space we expect the distribution of blogs to closely track proximity to surfing locations (primarily lakes) in Switzerland. Further, these users are more likely to visit those locations if wind speeds there are around or higher than 4 BFT. Hence, if wind speeds truly affect blogging, we would expect to see that blogs relating
to a particular location are more likely when wind speeds there are higher. We present geographic graphs that suggest that both aspects are true in the data.

Figure (1) plots the geographic distribution of blogging activity. From Figure (1) we see that many of the blogs emerge from Bern or Zurich, as many users are located there. Even so, more blogs emerge from areas closer to large lakes, especially in the north central part of the country. These are also the areas that see the most surfing activity in Switzerland. This correlation, while suggestive, is not conclusive, as the concentration of blogging could simply reflect the population density of users. Hence, we check whether the percentage of generated blogs relating to a particular surfing location is correlated with the wind speeds at that location. Figure (2) plots in red the percentage of total blogs written about a particular (lake) location, and in blue, the percentage of days where the wind at that location was greater than 4 BFT. We see that there is evidence of a strong correlation, implying that blogging responds to wind speeds at the focal location. Finally, we check whether users who generate content about a particular location are also those that have a large number of friends. This positive correlation is basic to the co-dependence of interest. Figure (3) adds user social connections to the first plot. The figure suggests evidence for a robust correlation, as more blogs about a location are generated by users who have more ties.

These plots evince patterns that suggest that blogging is linked to wind speeds, and that those who blog often also tend to be well connected. Our identification strategy is tied to using wind speeds as exogenous shifters of blogging, which are excluded from the propensity to have friends. The key assumption behind this exclusion argument is that users do not form friendships at surfing locations per se, so a higher wind speed does not directly cause a user to have more friends. We believe this exclusion assumption is reasonable as: (1) windsurfing is a not a team sport but one that is typically practiced alone, (2) verbal communication on the water is difficult due to wind conditions and the distances between the windsurfers, (3) windsurfing is only practiced when the wind exceeds 4 BFT, which happens only for short periods of time in Switzerland. Figure (3) provides evidence that individuals closely located to each other in geographic space are not frequently connected to each other on the website, as would be the case if users are primarily connecting with each other following their joint visits to local surf locations. The connectivity pattern in Figure (3) supports the anecdotal wisdom that users are primarily forming ties on the basis of their online interaction. Our informal conversations with the management of Soulrider.com support this pattern of behavior.

We now provide informal evidence for our assumptions – formal tests are provided in our estimation results. Denote $i$ for individual, $t$ for week, $b_{it}$ for blogs and $f_{it}$ for friends.
Variable | Obs | Mean | Std. Dev. | Min | Max |
---|---|---|---|---|---|
wind\_mean | 11776 | 2.32 | 0.44 | 1.5 | 3.57 |
windsd | 11776 | 0.58 | 0.24 | 0 | 1.67 |
winds max | 11776 | 3.17 | 0.71 | 1.5 | 6.00 |

Table 3: Descriptive Statistics of Wind Instruments

Let $wind\_mean_{it}$ denote the average daily wind speed at the local windsurfing location in week $t$ at individual $i$’s most preferred surfing location. Let $wind\_sd_{it}$ and $wind\_max_{it}$ respectively denote the standard deviation and maximum of wind speeds across days in week $t$ at individual $i$’s most preferred surfing location. Table (3) presents summary statistics of these variables. We collect in the vector $z_{it}$ all the wind variables and in $m_{it}$ all user-time varying shifters of friends (which are explained in Section (3.1.1)). Then we run a regression of friends, $f_{it}$, on blogs $b_{it}$, $m_{it}$ and $z_{it}$, controlling for individual and month fixed effects ($t$ statistics in parenthesis):

\[
 f_{it} = 0.046b_{it} + \alpha_i + \alpha_t + \beta_1m_{it} \\
 + 0.017wind\_mean_{it} + 0.012wind\_sd_{it} - 0.001wind\_max_{it}
\]

We find the $z_{it}$ variables are not significant in explaining friendship formation. While not formal, this provides an back-of-the-envelope assessment of the validity of the exclusion restriction.
Figure 1: Geographic Distribution of Blogging
Figure 2: Blogging Relationship to Wind Speeds
Figure 3: Blogs and Network Ties
3 Empirical Framework

We now discuss the empirical framework we adopt for the estimation of the model. We outline two different approaches, each entailing different assumptions or specifications, and show our results are robust to both. First, we outline a linear simultaneous equations model of content generation and friendship formation. We estimate the model by GMM. The GMM approach requires access to additional exogenous shifters of friends, in addition to the “wind” instruments for blogs. We discuss an identification strategy using the friendship requests received by the friends of each focal agent as instruments. In our second approach, we outline a nonlinear model, which takes into account the integer nature of the friends and blogs variables. We discuss estimation of these models using a GMM approach. As in the linear model, the nonlinear GMM approach requires additional instruments for the friends variable. In both approaches, we accommodate bi-level fixed effects to control for sources of spurious correlation in actions that could confound measures of causal effects.

3.1 Linear Model

We start with a linear simultaneous equations model linking blogs $b_{it}$ and friends $f_{it}$ where as before, $i$ stands for individual, and $t$ for week. As before, let the vector $z_{it} \equiv \{\text{wind\_mean}_{it}, \text{wind\_sd}_{it}, \text{wind\_max}_{it}\}$ denote the wind measures that affect blogs, but are excluded from the equation that determines the number of new friendship links requested. Let $m_{it}$ denote other variables that affect the number of friendship requests received by agent $i$, but which are excluded from the equation that determines his blogging activity (we discuss $m_{it}$ below).

The linear model is,

$$ b_{it} = \alpha_1 + \gamma_1 t + \theta_1 f_{it} + \delta_1 z_{it} + \varepsilon_{1it} \quad (1) $$
$$ f_{it} = \alpha_2 + \gamma_2 t + \theta_2 b_{it} + \delta_2 m_{it} + \varepsilon_{2it} \quad (2) $$

The causal effects of interest are $\theta = (\theta_1, \theta_2)$, the marginal effects of blogs and friends requests on each other. The individual-specific fixed effects in both equations control for time-invariant characteristics of individuals, and thus control for biases that may arise from homophily or endogenous group formation (e.g. Nair et. al. 2006). The time period fixed effects in both equations control for common sources of co-movement in friends and blogs and control for time varying unobservables that may generate spurious correlation.

---

4Endogenous group formation, or “homphily,” arises because agents with similar tastes may tend to form social groups; hence, subsequent correlation in their behavior may reflect these common tastes, and not a causal effect of one’s behavior on another. One solution to the endogeneity of group formation is facilitated by the availability of panel data. With panel data, one can control for endogenous group formation via agent fixed effects (e.g. Nair et al. 2006), or random effects that control for heterogeneity (e.g. Hartmann, 2008). Both fixed and random effects serve the role of picking up common aspects of group tastes.
Even after controlling for these, equation by equation estimation of this system by OLS is still inconsistent for $\theta$. To see this, consider Equation (2) for $f$. The RHS variable, $b_{it}$, is correlated with $\varepsilon_{2it}$ because (1), $b_{it} = b_{it} (\varepsilon_{2it})$ from Equation (1), and we do not a priori assume that $corr (\varepsilon_{1it}, \varepsilon_{2it}) = 0$, and (2) because $b_{it}$ is directly a function of $f_{it}$ by Equation (1). However, as long as $\delta_1 \neq 0$ and $\delta_2 \neq 0$, exclusion restrictions exist, and we can estimate the two parameters via GMM.

### 3.1.1 GMM Approach

We first discuss how we obtain the instruments $m_{it}$ for friends. The instruments are motivated by exclusion restrictions that exploit both the nature of the empirical context, as well as the structure of the social network. Suppose there are only 3 agents: the focal agent $i$, and agents $j$ and $l$. Suppose $i$ and $j$ are connected (i.e. already friends), but $l$ is not connected to either. Let $I_{l \rightarrow i,t}$ be an indicator of whether agent $l$ sends a friendship request to connect with the focal agent $i$ in period $t$. Then, the number of friend requests $i$ receives is $f_{it} = \sum_{l \rightarrow i,t}$. From equation (1) since $f_{it}$ affects $b_{it}$, $I_{l \rightarrow i,t}$ has a direct effect on $b_{it}$. Now suppose $l$ decides to send a friend request to $i$‘s friend $j$: i.e., $I_{l \rightarrow j,t} = 1$. Clearly, $I_{l \rightarrow j,t}$ is correlated with the endogenous variable $f_{it} = \sum_{l \rightarrow i,t}$ because factors that influenced $l$ to send a request to $i$ could also influence him to send a request to $i$‘s friend $j$. However, as $I_{l \rightarrow j,t}$ is not a request to $i$, $I_{l \rightarrow j,t}$ does not have a direct effect on $i$‘s blogging, $b_{it}$. Hence, $I_{l \rightarrow j,t}$ serves as an instrument for $f_{it}$. Essentially, we use the requests received by $i$‘s friends as instruments for the requests he receives. Intuitively, identification here derives from the structure of the network, as well as the fact that after controlling for any common shocks (via rich fixed-effects), an agent’s content generation is driven only by the environment facing him, and not his friends.

Figure (4) illustrates how the friends variable and the proposed instrument are operationalized. From the perspective of user $i$, the network may be visualized as a tree. Solid lines connecting nodes represent existing friendship links, while arrows represent friendship requests issued by one user to another. Placement of a user in one of the “levels” of the tree is determined by the minimum number of friendship links a user must “traverse” in order to reach user $i$ (i.e., the degree of separation). The friends variable for user $i$ is represented by the total number of dashed arrows pointing to node $i$. The proposed instrument for friends is the total number of friendship requests received by $i$‘s friends, or requests received at one degree of separation. This variable is given the name $fof$ (“friends of friends”) and corresponds to the dotted arrows in the figure. To reiterate the identification condition, we expect $fof$ to be positively correlated with friends through common characteristics which
dissipate as a function of the network distance. For example, immediate friends presumably have similar levels of “gregariousness”, which in turn leads to similar rates of friendship formation. Exclusion of fof from the blogs equation is based on the premise that user i does not set his blogging output in response to friendship requests received by anyone other than himself.

To formally define the fof_it variable, we first collect the set of friends i has in At(i). We then count the number of friend requests received by i’s friends in period t as fof_it = Σj∈At(i):j∉At(j)I_k→j, t. As with the wind data, observations of fof are of daily frequency whereas the panel frequency is weekly. We generate multiple instruments for friends as moments of the weekly aggregation of fof observations. That is, we compute mit = (mean(fof_id), SD(fof_id), max(fof_id)), where d indexes the daily observations corresponding to week t. Summary statistics of these variables are provided in Table (4).

Figure 4: Friendship network structure

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>fof_mean</td>
<td>11776</td>
<td>0.10</td>
<td>0.23</td>
<td>0</td>
<td>4.14</td>
</tr>
<tr>
<td>fof_sd</td>
<td>11776</td>
<td>0.16</td>
<td>0.34</td>
<td>0</td>
<td>3.62</td>
</tr>
<tr>
<td>fof_max</td>
<td>11776</td>
<td>0.40</td>
<td>0.86</td>
<td>0</td>
<td>11.00</td>
</tr>
</tbody>
</table>

Table 4: Descriptive Statistics of Friend of Friend Instruments
With these instruments, we can now base estimation on the moment conditions,

\[ M_1 = E[b_{it} - (\alpha_{1i} + \gamma_{1t} + \theta_1 f_{it} + \delta_1 z_{it}) \mid m_{it}] = 0 \]  
(3)

\[ M_2 = E[f_{it} - (\alpha_{2i} + \gamma_{2t} + \theta_2 b_{it} + \delta_2 m_{it}) \mid z_{it}] = 0 \]  
(4)

The parameters that jointly satisfy these moment conditions minimize the GMM objective function, \[ \begin{bmatrix} M_1 & M_2 \end{bmatrix} ^\prime \mathbb{W} \begin{bmatrix} M_1 & M_2 \end{bmatrix} \], where \( \mathbb{W} \) is a weighting matrix. Following Hansen (1982), the optimal \( \mathbb{W} \) is inversely proportional to the variance of the moments. We estimate \( \mathbb{W} \) via the usual 2-step procedure, assuming independent moments in the first step, and using the 1st-step estimate to construct the sample analog of the optimal weighting matrix in the second step.

### 3.1.2 Social Multiplier

Before discussing estimation of the nonlinear model, we use the linear model presented above to illustrate how a social multiplier arises in this setup. We briefly discuss how feedback from tie formation generates a multiplier for blogging when the website implements policies to subsidizing tie-formation. Suppose the website introduces an intervention that increases individual \( i \)'s friends by a small amount, \( \Delta \). Suppose, the cost to the website of this intervention is \( $c \). Consider a situation in which there is no feedback (i.e., only, Equation (1)). The incremental user-generated content produced by the intervention is \( \Delta \times \theta_1 \), and hence, the return on investment on the intervention, in terms of user-generated content is,

\[ ROI = \frac{\Delta}{c} \times \theta_1 \]  
(5)

Now consider the case where we incorporate the feedback from friend formation back into blogs. To see the full feedback effects, we can substitute Equation (2) into Equation (1) and obtain:

\[ b_{it} = \alpha_{1i} + \gamma_{1t} + \theta_1 \times [\alpha_{2i} + \gamma_{2t} + \theta_2 b_{it} + \varepsilon_{2it}] + \delta_1 z_{it} + \varepsilon_{1it} \]  
(6a)

\[ b_{it} = \frac{\alpha_{1i}}{1 - \theta_1 \theta_2} + \frac{\gamma_{1t}}{1 - \theta_1 \theta_2} + \frac{\theta_1}{\theta_1} \times [\alpha_{2i} + \gamma_{2t} + \varepsilon_{2it}] + \frac{\delta_1}{\delta} z_{it} + \tilde{\varepsilon}_{1it} \]  
(6b)

Now consider a small change in \( f \) in Equation (2). This affects blogs via Equation (1), which in turn affects friends via Equation (2) and so on till the system settles. The full effect of the intervention can be deduced from the reduced form of the system in Equation

---

5In the results section, we convert these into revenue terms by correlating user-generated content with page impressions, and advertising revenue. That is, we calculate the revenue from blogs as (Ad-dollars per page impression) \( \times \) (Page impressions per blog).
The incremental user-generated content produced by the intervention is $\Delta \times \frac{\theta_1}{1 - \theta_1 \theta_2}$, and hence, the ROI on the intervention is,

$$ROI_{\text{Multiplier}} = \frac{\Delta}{c} \times \frac{\theta_1}{1 - \theta_1 \theta_2}$$

(7)

This will be greater than $\theta_1$ as long as both effects are positive ($\theta_1 > 0, \theta_2 > 0$) and $\theta_1 \theta_2 < 1$ (we find both conditions are true in our empirical analysis). Moreover, the larger the effect of blogging on link formation (i.e., the larger the value of $\theta_2$), the larger will be the multiplier.

Hence, measuring the feedback is key to obtaining an accurate assessment of the ROI of marketing interventions on the site. It is also clear that estimating the causal effects $\theta_1$ and $\theta_2$ correctly is key to ROI measurement. Further, the reader should note that spurious correlation between blogs and friends, i.e. that fact that $corr(\varepsilon_{1it}, \varepsilon_{2it}) \neq 0$, does not imply any such multiplier on marketing effort. Only causal effects do.

### 3.2 A Nonlinear Model

We now discuss a nonlinear approach designed to accommodate potential nonlinear dependence between the blogs and friends variables. The empirical model is based on an exponential Poisson specification, which is consistent with the count nature of the data, but does not impose the integer specification explicitly. We estimate the model via GMM. The need to accommodate for user fixed effects complicates the estimation due to the large number of individual-specific parameters. We briefly present the approach without individual fixed effects below, and then discuss GMM estimation of the model with individual fixed effects in further detail.

#### No user fixed effects

We base GMM estimation of the parameters of the nonlinear model on conditional moment conditions on an exponential function of the mean (see Cameron and Trivedi 1998). This approach is consistent with a Poisson data generating process (i.e. the Poisson pseudo MLE leads to the same moment conditions), but does not impose the Poisson functional form. Moreover, estimation is based only the first moments of the data without imposing restrictions on the second or higher moments. This implies the GMM estimator does not impose the equidispersion property of the Poisson model. Estimation is based on the following moment conditions,

$$M_1 = \mathbb{E}[b_{it} - \exp(\alpha_1 + \gamma_{1it} + \theta_1 f_{it} + \delta_1 z_{it}) \mid m_{it}] = 0$$

(8)

$$M_2 = \mathbb{E}[f_{it} - \exp(\alpha_2 + \gamma_{2it} + \theta_2 b_{it} + \delta_2 m_{it}) \mid z_{it}] = 0$$

(9)

where, the $\alpha$-s are constant across users (i.e., no fixed effects). Although the moment conditions are now nonlinear functions, this poses little additional complication for GMM.
The remaining steps of the estimation procedure are the same as described above for the linear case.

With user fixed effects  We now discuss the estimation of the model when incorporating user fixed effects. The fixed effects result in a proliferation of parameters to be estimated, which is cumbersome in a nonlinear model. We discuss the procedure we adopt to concentrate out the fixed effects, which facilitates a nonlinear search of the GMM objective over the remaining parameters. Accommodation of the fixed effects requires us to take a stance on the distribution of the count variable. For the discussion below, we assume the distribution is Poisson. To motivate the moment conditions used in this case, consider the p.m.f. of $f_{it}$ assuming it is distributed Poisson,

$$
\Pr[f_{it} | \lambda_{2it}] = \frac{\exp(-\lambda_{2it}) \lambda_{2it}^{f_{it}}}{f_{it}!} \quad \text{and} \quad \lambda_{2it} = \exp(\alpha_{2i} + \gamma_{2t} + \theta_{2} b_{it} + \delta_{2} m_{it})
$$

Consider the log-likelihood of $f_{it}$,

$$
\mathcal{L}(\alpha_2, \gamma_2, \theta_2) = \sum_i \sum_t \left( -\exp(\alpha_{2i} + \gamma_{2t} + \theta_{2} b_{it} + \delta_{2} m_{it}) + f_{it} \times (\alpha_{2i} + \gamma_{2t} + \theta_{2} b_{it} + \delta_{2} m_{it}) - \ln(f_{it}!) \right)
$$

Setting the first order conditions of $\mathcal{L}(\alpha_2, \gamma_2, \theta_2)$ with respect to $\alpha_{2i}$ equal to 0 implies that,

$$
\sum_t \left( -\exp(\alpha_{2i} + \gamma_{2t} + \theta_{2} b_{it}) + f_{it} \right) = 0
$$

which implies $\exp(\alpha_{2i}) = \frac{\sum_t f_{it}}{\sum_t \exp(\gamma_{2t} + \theta_2 b_{it} + \delta_2 m_{it})} = \frac{f_{it}}{\lambda_{2it}}$ at the optimum.

These first order conditions imply the Poisson MLE is equivalent to a moment estimator in a model where the ratio of individual, or within group, means are used to estimate the individual fixed effects (Blundell et. al. 2002). Using this formulation, we now set up estimation of the parameters on the basis of the following moment conditions,

$$
\mathcal{M}_1 = \mathbb{E} \left[ b_{it} - \frac{b_i}{\lambda_{1i}} \exp(\gamma_{1t} + \delta z_{it} + \theta_1 f_{it}) \mid m_{it} \right] = 0 \quad (10)
$$

$$
\mathcal{M}_2 = \mathbb{E} \left[ f_{it} - \frac{f_i}{\lambda_{2i}} \exp(\gamma_{2t} + \theta_{2} b_{it} + \delta_{2} m_{it}) \mid z_{it} \right] = 0 \quad (11)
$$

The remaining steps of the estimation procedure are the same as described above for the case without user fixed effects.

3.3 Discussion

This section presented different approaches to estimate the joint system representing blogging and friendship formation. The estimators vary in the assumptions employed as well
as the approaches adopted, but are all designed to solve the fundamental issue that blogs and friends are codetermined. Codetermination complicates the estimation by generating simultaneity. We presented linear and nonlinear versions of the model to handle this co-dependence. Using the exogenous variation in wind-speeds, and under the null that actions targeted by friends of friends are valid instruments, the GMM estimator provides consistent estimates of all parameters. A rich set of fixed effects, facilitated by access to panel data, enables auxiliary controls for other spurious sources of correlation, which have typically plagued empirical models of social interactions.

4 Results

We now discuss the results from the estimation of the above models on the Soulrider data. We present the results in the following sequence. First, we present results from the linear model. For the linear models, we present results for OLS, with and without fixed effects to illustrate the importance of endogenous group formation. We then present results from the GMM estimation for the linear model. Subsequently, we present results from the nonlinear count model.

4.1 Linear Model: OLS

Table (5) presents results from OLS regressions of friends on blogs and of blogs on friends. Looking at Table (5), we see there is preliminary evidence of a feedback effect. Both the effect of friendship links on blogs, and of blogs on links are strongly statistically significant. However, note these estimates are likely upward biased due to the lack of control for endogenous group formation. We also see that the wind variables are significant in explaining blogging. The table also presents results from including month fixed effects to control for time-varying unobservables that drive blogging and friendship formation. We see that month fixed effects do not change the estimates that much, suggesting that such common unobservables are not first-order for these data. Nevertheless, the direction of the change in the estimate is consistent with our intuition: once we control for this spurious source of correlation, we expect the parameters on blogs/friends respectively to decrease in magnitude. Referring to Table (5), we see that this is indeed the case.

We now discuss the results from adding individual fixed effects into the previous specification. The addition of individual fixed effects is facilitated by the availability of panel data. The reader should note that this specification is very demanding of the data as the inclusion of both individual and month fixed effects imply that all variation common to individuals within a given month, as well as variation common to months for a given indi-
vidual, are fully controlled for, and are not used to inform the causal effects of blogs and friends on each other. The fixed effects control for endogenous group formation, and are expected to correct an upward bias in the estimation of the causal effects. Looking at the last column of Table (5), we see that this is indeed the case. Looking at column 3, we see that the effect of blogs on friends has dropped from 0.068 to 0.051 when adding user fixed effects (a 25% decrease). The effect of friends on the other hand has dropped from 0.251 to 0.174 (a 30% decrease). We see that controlling for endogenous group formation as well as common unobservables is important for these data especially for the effect of friendship on the generation of blogging. We also see that individual fixed effects also control for a large source of unobserved within-user persistence in the data. For testing statistical significance, note that all tables report robust standard errors reported that have been clustered at the user level.

4.2 Linear Model: GMM Estimation

Table (6) now presents estimates of the linear model when instrumenting for friends and blogs using GMM. We first discuss whether these instruments are working correctly. First, we discuss whether we are subject to a weak instruments problem. To test for weak instruments, we report the Kleibergen-Paap (2006) \( rk \) statistic, which is a generalization of the weak IV test to the case of non-i.i.d. errors. The null is that the instruments are weak, and a rough thumb rule for empirical work is that there is no weak instruments problem if the \( rk \) statistic is > 10. Looking at the columns in Table (6), we see that this is the case: the weak identification statistics are all > 20. Further, we see that the overidentifying restrictions for the instruments are not rejected in any of the models, and that the fit is good. Overall, these diagnostics indicate that the instruments are working properly.

We also see that the magnitude of the coefficient on \textit{blogs} in the \textit{friends} equation has increased after instrumenting. This is plausible, and is consistent with stories where the unobservables are negatively correlated with the endogenous variable. We do not take a strong stance on what these unobservables represent. For instance, we conjecture that unobservables that drive friendship formation could proxy for extroversion, friendliness or windsurfing skill (more users want to be friends with better windsurfers all things held equal). If extroverts post more blogs, the unobservables would be positively correlated with blogs, and we would observe an upward bias. If on the other hand, extroverts tend to spend more time offline, and post fewer blogs, then unobservables would be negatively correlated with blogging, and we would observe a downward bias (as we see here). A priori it either story seems reasonable.
<table>
<thead>
<tr>
<th>Regressions of b(f):</th>
<th>OLS</th>
<th>OLS</th>
<th>OLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$</td>
<td>0.251***</td>
<td>0.228***</td>
<td>0.174***</td>
</tr>
<tr>
<td>(0.043)</td>
<td>(0.041)</td>
<td>(0.027)</td>
<td></td>
</tr>
<tr>
<td>wind_mean</td>
<td>0.106***</td>
<td>0.178***</td>
<td></td>
</tr>
<tr>
<td>(0.027)</td>
<td>(0.025)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wind_sd</td>
<td>0.236***</td>
<td>0.244***</td>
<td></td>
</tr>
<tr>
<td>(0.051)</td>
<td>(0.049)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wind_max</td>
<td>-0.040</td>
<td>-0.056**</td>
<td></td>
</tr>
<tr>
<td>(0.022)</td>
<td>(0.021)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>constant</td>
<td>0.109***</td>
<td>-0.226***</td>
<td></td>
</tr>
<tr>
<td>(0.010)</td>
<td>(0.038)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Month effects | No | Yes | Yes |
| Individual effects | No | No | Yes |
| Obs            | 11776 | 11776 | 11776 |
| Rsq            | 0.02 | 0.05 | 0.05 |
| RMSE           | 0.453 | 0.446 | 0.397 |

<table>
<thead>
<tr>
<th>Regressions of f(b):</th>
<th>OLS</th>
<th>OLS</th>
<th>OLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b$</td>
<td>0.068***</td>
<td>0.046***</td>
<td>0.051***</td>
</tr>
<tr>
<td>(0.013)</td>
<td>(0.011)</td>
<td>(0.009)</td>
<td></td>
</tr>
<tr>
<td>fof_mean</td>
<td>0.029</td>
<td>-0.021</td>
<td></td>
</tr>
<tr>
<td>(0.050)</td>
<td>(0.046)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fof_sd</td>
<td>0.055</td>
<td>0.187</td>
<td></td>
</tr>
<tr>
<td>(0.086)</td>
<td>(0.104)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fof_max</td>
<td>0.028</td>
<td>-0.007</td>
<td></td>
</tr>
<tr>
<td>(0.035)</td>
<td>(0.039)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>constant</td>
<td>0.039***</td>
<td>0.013**</td>
<td></td>
</tr>
<tr>
<td>(0.003)</td>
<td>(0.004)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Month effects | No | Yes | Yes |
| Individual effects | No | No | Yes |
| Obs            | 11776 | 11776 | 11776 |
| Rsq            | 0.02 | 0.06 | 0.04 |
| RMSE           | 0.236 | 0.231 | 0.223 |

| p value levels | * p<0.05 | ** p<0.01 | *** p<0.001 |

Table 5: Linear Model: OLS
## Regressions of $b(f)$

<table>
<thead>
<tr>
<th></th>
<th>IV1</th>
<th>IV1</th>
<th>IV1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$</td>
<td>1.690***</td>
<td>1.559***</td>
<td>0.854***</td>
</tr>
<tr>
<td></td>
<td>(0.278)</td>
<td>(0.284)</td>
<td>(0.171)</td>
</tr>
<tr>
<td>wind_mean</td>
<td>0.066*</td>
<td>0.151***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.033)</td>
<td>(0.026)</td>
<td></td>
</tr>
<tr>
<td>wind_sd</td>
<td>0.175**</td>
<td>0.221***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.060)</td>
<td>(0.051)</td>
<td></td>
</tr>
<tr>
<td>wind_max</td>
<td>-0.020</td>
<td>-0.048*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.022)</td>
<td></td>
</tr>
<tr>
<td>constant</td>
<td>0.041***</td>
<td>-0.196***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.043)</td>
<td></td>
</tr>
</tbody>
</table>

- Month effects: No, Yes, Yes
- Individual effects: No, No, Yes
- Obs: 11776, 11776, 11776
- RMSE: 0.568, 0.547, 0.433
- Weak id F: 27.063, 26.331, 30.197
- Over id Chi Sq: 4.056, 4.174, 5.838
- Over id dof: 2, 2, 2
- Over id pval: 0.132, 0.124, 0.054

## Regressions of $f(b)$

<table>
<thead>
<tr>
<th></th>
<th>IV2</th>
<th>IV2</th>
<th>IV2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b$</td>
<td>0.189***</td>
<td>0.116*</td>
<td>0.150***</td>
</tr>
<tr>
<td></td>
<td>(0.051)</td>
<td>(0.054)</td>
<td>(0.039)</td>
</tr>
<tr>
<td>fof_mean</td>
<td>0.018</td>
<td>-0.045</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.048)</td>
<td>(0.048)</td>
<td></td>
</tr>
<tr>
<td>fof_sd</td>
<td>0.027</td>
<td>0.178</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.092)</td>
<td>(0.106)</td>
<td></td>
</tr>
<tr>
<td>fof_max</td>
<td>0.035</td>
<td>-0.004</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td>(0.039)</td>
<td></td>
</tr>
<tr>
<td>constant</td>
<td>0.024***</td>
<td>0.013**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.004)</td>
<td></td>
</tr>
</tbody>
</table>

- Month effects: No, Yes, Yes
- Individual effects: No, No, Yes
- Obs: 11776, 11776, 11776
- RMSE: 0.243, 0.233, 0.230
- Weak id F: 32.973, 26.151, 37.938
- Over id Chi Sq: 0.517, 0.091, 0.304
- Over id dof: 2, 2, 2
- Over id pval: 0.772, 0.956, 0.859

### Table 6: Results from Linear Model: GMM Estimation
4.3 Nonlinear Models

We now present estimates for nonlinear count models. We expect these models to perform better than the linear model as they take into account potential nonlinear dependence between blogging and friendship requests (and hence may produce more efficient estimates). We present the results in Table (7). As a benchmark, we start by presenting GMM estimates of friends and blogs, in which we ignore the endogeneity of either variable (that is, we do not instrument for either). In this model, we control for user and month fixed effects. These results are presented in column 1 of Table (7). We see that the effect of blogs and friends are strongly significant in both models. Columns 2-4 of Table (7) present results from GMM estimation of the model in which we instrument for friends and blogs as before, but vary the fixed effects specification. Looking at these results, we see the basic pattern of the results suggest a feedback effect: blogs have a significant effect on friends and the other way around. Further, the overidentifying restrictions implied by the moments are not rejected. Column (4) of Table (7) presents the results from our preferred specification with a full set of bi-level fixed effects.

5 Implication for ROI

We demonstrate the application of our results by considering the return on investment (ROI) from subsidizing link formation. The counterfactual policy of interest adds $\Delta$ to the number of friendship requests received by each user (the friends variable) at a cost of $c$ monetary units per user. Such a policy might be implemented, for example, by developing a “friend suggestion” application that (in expectation) yields $\Delta$ incremental friendship requests. Without loss of generality, we may set $\Delta = 1$ and interpret $c$ as the cost per user to subsidize formation of one additional friendship link. As the value of $c$ will vary depending on the mechanism chosen to implement the policy, we focus on comparing the expected return on the implementation of the policy to an exogenously specified hurdle rate. That is, we combine the model estimates with revenue data to determine an upper bound on $c$ that will achieve a target level of return. To quantify the contribution of feedback effects to the total return, we also compare the return incorporating feedback with a baseline model in which the feedback effects are ignored. This difference captures the extent to which the site operator would underestimate the value of the policy were he to ignore the simultaneity of friends and blogs.

As with most social networking sites, Soulrider.com’s primary revenue stream is from online advertising. Soulrider.com receives approximately 0.002 Swiss Francs (CHF) per
### Regressions of $b(f)$:

<table>
<thead>
<tr>
<th></th>
<th>GMM (no IV)</th>
<th>GMM IV1</th>
<th>GMM IV1</th>
<th>GMM IV1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$</td>
<td>0.502***</td>
<td>2.116***</td>
<td>1.969***</td>
<td>0.962***</td>
</tr>
<tr>
<td></td>
<td>(0.093)</td>
<td>(0.160)</td>
<td>(0.206)</td>
<td>(0.221)</td>
</tr>
<tr>
<td>wind_mean</td>
<td>1.121***</td>
<td></td>
<td>0.167</td>
<td>1.023***</td>
</tr>
<tr>
<td></td>
<td>(0.169)</td>
<td></td>
<td>(0.312)</td>
<td>(0.167)</td>
</tr>
<tr>
<td>wind_sd</td>
<td>0.767*</td>
<td>0.777</td>
<td>0.685</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.356)</td>
<td>(0.312)</td>
<td>(0.378)</td>
<td></td>
</tr>
<tr>
<td>wind_max</td>
<td>-0.035</td>
<td>0.213</td>
<td>-0.008</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.154)</td>
<td>(0.415)</td>
<td>(0.164)</td>
<td></td>
</tr>
<tr>
<td>constant</td>
<td>-2.886***</td>
<td>-5.301***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.194)</td>
<td>(0.702)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.392</td>
<td>2.586</td>
<td>1.300</td>
<td>0.403</td>
</tr>
<tr>
<td>Over id Chi Sq</td>
<td>0.201</td>
<td>0.252</td>
<td>1.788</td>
<td></td>
</tr>
<tr>
<td>Over id p-value (dof)</td>
<td>0.905 (2)</td>
<td>0.882 (2)</td>
<td>0.409 (2)</td>
<td></td>
</tr>
</tbody>
</table>

### Regressions of $f(b)$:

<table>
<thead>
<tr>
<th></th>
<th>GMM (no IV)</th>
<th>GMM IV2</th>
<th>GMM IV2</th>
<th>GMM IV2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b$</td>
<td>0.372***</td>
<td>1.034***</td>
<td>0.764***</td>
<td>0.973***</td>
</tr>
<tr>
<td></td>
<td>(0.096)</td>
<td>(0.105)</td>
<td>(0.148)</td>
<td>(0.270)</td>
</tr>
<tr>
<td>fof_mean</td>
<td>-0.033</td>
<td>-0.076</td>
<td>-0.196</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.272)</td>
<td>(0.248)</td>
<td>(0.534)</td>
<td></td>
</tr>
<tr>
<td>fof_sd</td>
<td>2.615***</td>
<td>1.839**</td>
<td>2.901***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.572)</td>
<td>(0.582)</td>
<td>(0.828)</td>
<td></td>
</tr>
<tr>
<td>fof_max</td>
<td>-0.585**</td>
<td>-0.302</td>
<td>-0.664**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.195)</td>
<td>(0.203)</td>
<td>(0.244)</td>
<td></td>
</tr>
<tr>
<td>constant</td>
<td>-3.516***</td>
<td>-3.909***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.134)</td>
<td>(0.166)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.223</td>
<td>0.302</td>
<td>0.248</td>
<td>0.223</td>
</tr>
<tr>
<td>Over id Chi Sq</td>
<td>0.094</td>
<td>0.907</td>
<td>0.515</td>
<td></td>
</tr>
<tr>
<td>Over id p-value (dof)</td>
<td>0.954 (2)</td>
<td>0.636 (2)</td>
<td>0.773 (2)</td>
<td></td>
</tr>
</tbody>
</table>

- **Month effects**: Yes No Yes Yes
- **Individual effects**: Yes No No Yes
- **Obs**: 11776 11776 11776 11776

1. Instruments for $f$: fof_mean, fof_sd, fof_max
2. Instruments for $b$: wind_mean, wind_sd, wind_max

Table 7: Results from Nonlinear Model
page impression (PI). To estimate the relationship between blog production and page impressions, we regress the blogs variable on the number of page-impressions on the user’s webpage (t statistics in parenthesis):

\[ PI_{it} = a_i + Xb_{it} = a_i + 368.04b_{it} \]

Thus, we assume a blog revenue function in Swiss Francs of \( R(b) = 0.002 \times 368.0 \times b = 0.736 \times b \).

Calculation of the contribution of feedback effects to the return requires solving the reduced form of the model, and is therefore specification-dependent. The calculation for the linear model was discussed in Section (3.1.2). Apart from converting to currency units, the treatment here is the same. Referencing our preferred linear model specification, which instruments for endogenous variables and incorporates both individual and period fixed effects, we obtain the following ROI measure:

\[
\text{ROI Multiplier} = \frac{0.736}{c} \times \frac{\hat{\theta}_1}{1 - \hat{\theta}_1 \hat{\theta}_2} = \frac{0.736}{c} \times \frac{0.854}{1 - 0.854 \times 0.150} = \frac{0.721}{c}
\]

This calculation implies, for example, that for the policy to clear a hurdle rate of 10%, the cost of the policy per user must be \( c \leq 7.2 \text{ CHF} \). If feedback effects are ignored, we obtain an ROI of:

\[
\text{ROI} = \frac{0.736}{c} \times \hat{\theta}_1 = \frac{0.736}{c} \times 0.854 = \frac{0.629}{c}
\]

The contribution of feedback effects to ROI is therefore \( \left( \frac{0.721 - 0.629}{0.629} \right) = 14.6\% \), which is an economically material increase.

In the case of our non-linear model, calculation of ROI is more involved. The complication arises due to the fact that solving the reduced form of the model involves finding the fixed point of the system,

\[
\begin{align*}
    b_{it} &= \exp (g(f_{it})) + \varepsilon_{1it} \quad (12) \\
    f_{it} &= \exp (h(b_{it})) + \varepsilon_{2it} \quad (13)
\end{align*}
\]

where \( g(\cdot) \) and \( h(\cdot) \) are linear functions. Since existence of the fixed point is not guaranteed at all points in the domain, we pursue an alternative strategy to evaluate the feedback effect. Rather than solve the nonlinear system exactly, we solve a linearized approximation of the model that holds in the vicinity of the user-specific mean values of \( g(\cdot) \) and \( h(\cdot) \). To obtain the approximation, we compute the first order Taylor expansions of \( b \) and \( f \) about the

\(^6\)1 CHF is roughly = 1 USD.
mean values and then solve the resulting linear reduced form. For brevity, we introduce the following shorthand notations for the mean values: 
\[ \bar{u}_1 = \alpha_{1i} + \gamma_1 + \delta_1 \bar{z}_i + \theta_1 \bar{f}_i \]
and
\[ \bar{u}_2 = \alpha_{2i} + \gamma_2 + \delta_2 \bar{m}_i + \theta_2 \bar{b}_i. \]
Then, the expansion about \((\bar{u}_1, \bar{u}_2)\) may be written as,
\begin{align*}
    b_{it} &\approx \exp(\bar{u}_1) (1 + \alpha_{1i} + \gamma_1 + \delta_1 z_{it} + \theta_1 f_{it} - \bar{u}_1) + \varepsilon_{1it} \tag{14} \\
    f_{it} &\approx \exp(\bar{u}_2) (1 + \alpha_{2i} + \gamma_2 + \delta_2 m_{it} + \theta_2 b_{it} - \bar{u}_2) + \varepsilon_{2it} \tag{15}
\end{align*}

With a linear approximation to the system in hand, we may now proceed in solving the reduced form as before:
\begin{align*}
    b_{it} &= \exp((\bar{u}_1)) \times \{ 1 + \alpha_{1i} + \gamma_1 + \delta_1 z_{it} \\
    &+ \theta_1 \times (\exp((\bar{u}_2)) (1 + \alpha_{2i} + \gamma_2 + \delta_2 m_{it} + \theta_2 b_{it} - \bar{u}_2) + \varepsilon_{2it} - \bar{u}_1) \} + \varepsilon_{1it} \\
    f_{it} &= \frac{\exp((\bar{u}_1))}{1 - \exp((\bar{u}_1)) \exp((\bar{u}_2)) \theta_1 \theta_2} \times \{ 1 + \alpha_{1i} + \gamma_1 + \delta_1 z_{it} \\
    &+ \theta_1 \times (\exp((\bar{u}_2)) (1 + \alpha_{2i} + \gamma_2 + \delta_2 m_{it} + \theta_2 b_{it} - \bar{u}_2) + \varepsilon_{2it} - \bar{u}_1) \} + \varepsilon_{1it}
\end{align*}

We can simplify as,
\begin{align*}
    b_{it} &= \frac{\exp((\bar{u}_1))}{1 - \exp((\bar{u}_1)) \exp((\bar{u}_2)) \theta_1 \theta_2} \times \{ 1 + \alpha_{1i} + \gamma_1 + \delta_1 z_{it} \\
    &+ \theta_1 \times (\exp((\bar{u}_2)) (1 + \alpha_{2i} + \gamma_2 + \delta_2 m_{it} + \theta_2 b_{it} - \bar{u}_2) + \varepsilon_{2it} - \bar{u}_1) \} + \varepsilon_{1it} \\
    f_{it} &= \frac{\exp((\bar{u}_1))}{1 - \exp((\bar{u}_1)) \exp((\bar{u}_2)) \theta_1 \theta_2} \times \{ 1 + \alpha_{1i} + \gamma_1 + \delta_1 z_{it} \\
    &+ \theta_1 \times (\exp((\bar{u}_2)) (1 + \alpha_{2i} + \gamma_2 + \delta_2 m_{it} + \theta_2 b_{it} - \bar{u}_2) + \varepsilon_{2it} - \bar{u}_1) \} + \varepsilon_{1it}
\end{align*}

Therefore, incorporating feedback effects, ROI in the case of our preferred nonlinear model will be:
\begin{align*}
    ROI_{Multiplier} = \frac{R(b(f + \Delta)) - R(b(f))}{c} \\
    &= 0.736 \times \frac{\hat{\theta}_1 \exp(\bar{u}_1)}{1 - \exp((\bar{u}_1)) \exp((\bar{u}_2)) \theta_1 \theta_2} \tag{16}
\end{align*}

Analogously, the multiplier with no feedback will be:
\begin{align*}
    ROI = 0.736 \times \frac{\hat{\theta}_1 \exp(\bar{u}_1)}{c} \tag{17}
\end{align*}

Note that the nonlinear approximation implies that ROI will differ across users due to the presence of the \(\exp((\bar{u}_i))\) terms in equations (12) and (13) above. As a consequence, the nonlinear model provides some measure of variation in ROI due to unobserved heterogeneity in user characteristics. To obtain a point estimate for the ROI multiplier (ROI with the \(\frac{1}{c}\) term factored out), we compute the multiplier on an individual by individual basis and report the sample average. These ROI multipliers are summarized in Table (8) below, where table columns correspond to the nonlinear model under the fixed effect specifications previously considered. The reported multipliers are reasonably consistent across the various models. Focusing on our preferred specification (last column), the estimates imply that for the policy to clear a hurdle rate of 10%, the cost of the policy per user must be \(c \leq 1.4\).
The contribution of feedback effects to ROI is $\frac{0.135 - 0.123}{0.123} = 10.2\%$. The nonlinear specification is therefore more conservative in its evaluation of the profitability of subsidizing tie formation and the fraction of returns due to feedback effects.

<table>
<thead>
<tr>
<th>Month effects</th>
<th>No</th>
<th>Yes</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual effects</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Mean ROI multiplier (no feedback)</td>
<td>0.129</td>
<td>0.158</td>
<td>0.123</td>
</tr>
<tr>
<td>Mean ROI multiplier (with feedback)</td>
<td>0.147</td>
<td>0.169</td>
<td>0.135</td>
</tr>
<tr>
<td>Feedback % increase</td>
<td>14.3</td>
<td>7.0</td>
<td>10.2</td>
</tr>
</tbody>
</table>

Table 8: Return on investment under the nonlinear model

To further explore the effects of user heterogeneity on ROI, we provide histograms of the ROI multipliers. Figure (5) below characterizes the distribution of ROI multipliers without feedback effects in our sample. This distribution is highly skewed, demonstrating that the proposed policy has moderate impact on average users but can have very significant impact for users with a high native propensity to blog. The large differences accentuate the need for the correct identification of subsets of users who respond to targeting, and the correct measurement of the ROI from interventions that induce them to post more content. Figure (6) is a histogram of the percentage increase to returns once feedback effects are incorporated. The presence of several high return individuals in the sample suggests that a targeted policy intervention for these users may deliver far higher rates of return than a uniform policy applied to all users. For example, one could heavily promote individuals with high ROI multipliers as potential friends for others via a “friendship recommendation” feature, presumably at very low cost. The approach presented here thus serves to both identify segments of users for whom targeted interventions matter most, as well as provides a clearer picture of the return on investment from these interventions.

6 Conclusions

We identify the role of social ties in generating content on online social networks. This paper adds to an emerging literature relevant to social network firm’s monetization strategies. We show that policies aimed at enhancing social ties can indirectly help facilitate content generation on the site, thereby creating a link between social tie-formation and advertising revenues. The indirect approach to content management is particularly relevant for network growth, as direct interventions to induce users to post content are either ineffective, or have proven harmful to growth.

Our approach develops ways to measure the return on investment of subsidizing tie
formation as a way to indirectly increase content generation on the network. We recognize that feedback effects between content generation and tie formation, if strong, have the potential to improve the ROI profile on these investments. We use rich detailed data from an online social network to conduct our empirical analysis. The data comprise the complete details of social tie formation and content generation on the site. The richness of the data enable us to control for several spurious confounds that have typically plagued empirical analysis of social interactions. The main challenge in the analysis is to separately identify the causal effect of social ties and content on each other, and to separate true causal effects from spurious sources of correlation. We use variation in wind speeds at surfing locations in Switzerland as exogenous shifters of users’ propensity to post content about their surfing activity. Our results show evidence for significant positive feedback in user generated content, and accentuate the importance of controlling for sources of spurious correlation in behavior in order to obtain a true picture of the ROI from targeted marketing interventions. As an additional by-product, our approach also enables us to identify, in an ex-post sense, the set of users that respond most to potential interventions. We discuss the implications of the estimates for the management of the content and the growth of the network.

Figure 5: Distribution of ROI multipliers without feedback effects
Figure 6: Percentage increases in ROI from feedback effects

7 References


