NET Institute*

www.NETinst.org

Working Paper #09-28

November 2009

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

Klaus Miller
University of Bern

* 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[‡] Klaus Miller[§]

November 2009[¶]
Net Institute Working Paper
WORK-IN-PROGRESS. DO NOT CITE WITHOUT PERMISSION.

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 conent generation in turn has a similar causal effect on the users' abilty 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 faciliate 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 details of social tie formation and content generation on the site. The richness of he data enable us to control for several spurious confounds that have typically plagued empirical analysis of social interactions. Our results show evidence for significant positive feedback in user generated content. 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 Student in Marketing, Stanford GSB, Email: scott.shriver@gsb.stanford.edu

[‡] Associate Professor of Marketing, Stanford GSB, Email: harikesh.nair@stanford.edu.

[§] Assistant Professor of Marketing, University of Bern.

[¶] 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 network sites. 75% of social network site users logged in at least once a week and 57% did so daily. IDC also found more than 61% of those users spent more than 30 minutes per session on social network sites, and 38% remained active for 1 hour or more (eMarketer 2008a). Despite their rising importance, the extent to which user behavior on online social networks has been systematically investigated in academia and practice is limited (e.g. Trusov et. al. 2009; *placeholder for other citations*).

A key challenge facing social network 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 largely reflect their personal activities and tastes, and may not be significantly influenced by features the site can control, like 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, 10.22% of the users account for 80% of the generated content). One option for the website is to selectively provide higher access to tools for these groups of users. However, installed-base competition amongst social networks has in practice typically precluded discriminating amongst users in their access to tools (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, directly influencing users to generate content is also not very feasible. For instance, paying users to generate content is especially tricky, as it biases the character of the content, which can have negative repercussions in some situations (e.g. the controversy over WalMart paying Edelman PR to create positive blog listings for its products; see Glaser 2006; and for more formal demonstration see, Friestad and Wright, 1994 and Verlegh et al. 2004). In essence, basic conundrum for social media is that when the site tries to influence the user-generated space, it loses its appeal for the user who wants to be in control.

We consider a separate mechanism by which the firm may indirectly manage the generation of content on its site. Our main insight is driven by the fact that the motivation of users to post content is derived by 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 the content generation is a function of the extent of social ties the user enjoys on the website. Thus, an indirect way to manage content for the firm is to facilitate social-tie formation, thereby influencing the users' network structure. If 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 policies that encourage regular information provision to users about the tastes, profiles and activities of others on the website which facilitate link formation with others. Here, we recognize that this effort has an additional benefit by providing an indirect way for the site to facilitate content generation.

The goals of this paper are three-fold. First, we test whether network structure has a causal effect on the generation of user content on a social network. Testing this is not straightforward, as unobservables that drive content generation may also drive link formation, thereby generating spurious correlation between both. The main identification challenges are endogenous group formation, correlated unobservables and simultaneity (e.g. see Manski 1993; Moffit 2001; and Nair et. al. 2006 for instance). We discuss how we overcome these challenges in our analysis. Second, we ask whether there exists a reverse causal effect whereby tie formation is driven by content generation itself. This is plausible as those that generate more content may be invited to network more with others. We are interested in assessing the reverse effect because of the potential for *multipliers* (e.g. Nair et. al. 2006). In particular, if both causal effects are strong, the interaction between content and network structure generates a social multiplier that accentuates the benefits to content generation from subsidizing tie-formation. Separating the other sources of correlations in observed behavior from true causal effects is key to measuring these multipliers. Only causal effects can result in a social feedback. Hence, uncovering causal effects accurately is key to formulating policy. 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 effects of such feedback effects on the return on investment from marketing on the network side.

We address these research questions using rich, detailed data from an online social network based in Switzerland, named Sourlrider.com. Soulrider is one of the largest sports-based community in Switzerland, and is primarily focused on windsurfers. Users post content on Soulrider about their surfing activities, as well as blogs including the information available to them about wind speeds and surf conditions at specific surfing locations. Our data comprises the complete details 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 enables us to use within-member variation in the joint distribution of social networks 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 detailed 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.[1] 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 reveals that evidence for significant causal effects in both directions. We find that the number of a user's social ties has a positive effect on his content generation, and that content generation in turn has a significant positive effect on his social ties. Our results are robust across a variety of estimators that transparently incorporate the sources of identification discussed above. 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. In ongoing work, we are extending the models in the paper to further explore the implications of our estimates for the management of the network.

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 5 presents the results, and the last section concludes.

[1] BFT stands for "Beaufort," the international wind scale used in weather reporting.

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. These users originate mainly from European countries and tend to be avid enthusiasts of extreme sports such as windsurfing (46.1%), snowboarding (27.4%), mountain biking (14.2%), or others. Registered users learn about Soulrider.com through word-of-mouth (WOM) (37.35%), a search engine (29.62%), an external link (16.66%), an invitation E-mail (5.66%), or other reasons (10.70%).

Users of Soulrider.com can consume both existing content or submit their own. Most content on the website, such as blogs, or forum messages, is generated by users themselves. Other content, such as 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 activites are required to create a free account on the website.

Besides 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, groups of interest, and a mandatory "add as friend" function, which handles the mechanics of creating ties in the online interface. In addition, 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 (11.70 PI-s per visit) 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. New 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 web-site for 4 minutes on average and generate approximately 400,000 page impressions per month (information from Google Analytics as of December 20th, 2008, for the past 30 days). 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 specific, 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 windsurfers on the website. Thus, our panel is of size 11,776 (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 (a date, surf-location, or other information), or photo(s). Thus, if a user posted a photo and a text message on the website, or just two text messages, or just two photos, blogs = 2. We do not account for the volume or number of photos added, focusing only on the *incidence* of postings, and not on its magnitude or type. In ongoing work, we are analyzing these alternative measures as well. We also work with a simple representation of social structure, whereby, we create a variable, *friends*, which counts the number of declared friends for each user. In other results (not reported), we have experimented with different measures of network position like centrality, and found the broad nature of our results remains consistent with the results using the *friends* variable.

2.2.1 Basic patterns in the data

We now discuss some 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 for supporting the identification provided by exogenous wind forecast data on content generation.

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 0.1 friends per week (maximum 11). There are also a large number of user-weeks when no blogs are posted, or no friends are added.

To see the spread visually, Figures (1) shows a histogram of friends attained by each user per week, and Figure (2) shows a histogram of blogs posted by each user per week. The distribution of both is highly skewed, will large mass point at zero. The histograms

Variable	Obs	Mean	Std. Dev.	Min	Max
blogs	11776	0.12	0.457	0	6.00
friends	11776	0.10	0.457	0	11.00

Table 1: Descriptive Statistics of Blogs and Friends

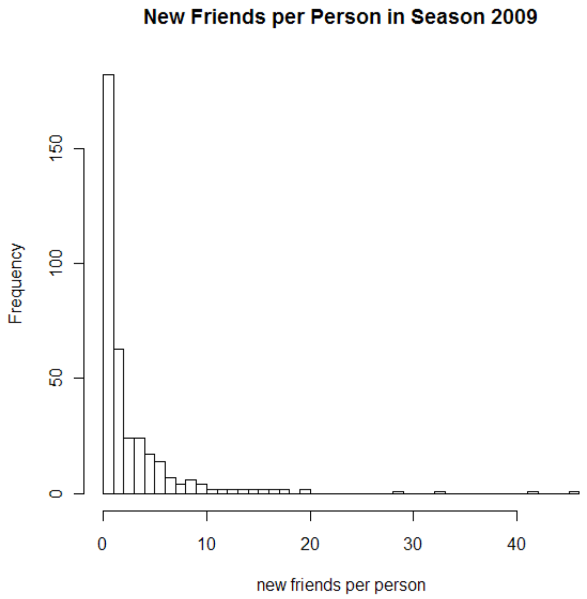

Figure 1: Histogram of *friends* variable

suggest 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 surfing location. In particular, users that are close to a viable surfing location are likely to visit those locations often, prompting them to blog about these more 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 emanating about 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 (3) plots the geographic distribution of blogging. From Figure (3) we see that

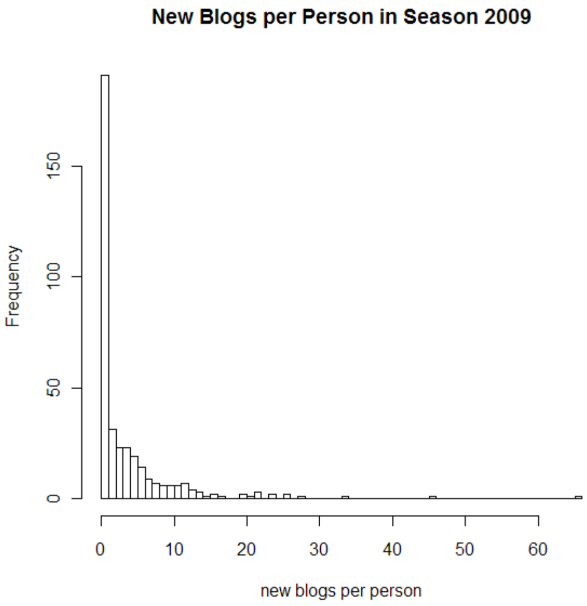

Figure 2: Histogram of *blogs* variable

many of the blogs emerge from Bern or Zurich, as many users are located there. Even so, more blogs seem to 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 blogging could reflect the location of users *per se*. Hence, we now check whether of the generated blogs, the percentage emanating about a particular surfing location is correlated with the wind speeds at that location. Figure (5) 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 wind speeds do significantly shift blogging.

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 we are inetrested in. Figure (4) adds the social connections of the users into the previous plot. The figure suggests evidence for a robust correlation. More blogs about a location are generated by users who have more ties.

To summarize these plots, we see broad 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 crosses 4 BFT, and in Switzerland, winds often stay above this threshold only for short periods of time. Hence, the timing of surfing is limited; while these are heuristic arguments, the strongest case comes from Figure (4) which shows the social connections of the users. Here, we see that most of the links connect users dispersed widely across the country. We do not see patterns where individuals closely located to each other in geographic space (for example, in the same city) are connected to each other, which would likely be the case if users are primarily connecting with each other following their joint visits to local surf location. The connectivity pattern in Figure (4) 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 confirms this pattern of behavior.

We now report on some support for this assumption in the data. Denote i for individual, t for week, b_{it} for blogs and f_{it} for friends. Let w_{it} denote the proportion of days in week t when the wind speed at individual i's most preferred surf-location was greater than 4 or equal to BFT. Let s_{it} and m_{it} respectively denote the mean and standard deviation of wind-speeds across days in week t at individual i's most preferred surf-location. Table (2) presents summary statistics of these variables. We see that wind speeds were condusive to surfing roughly two days on average per week during the 2009 season. Collecting these instruments in a vector $z_{it} = (w_{it}, s_{it}, m_{it})$, we run a regression of friends, f_{it}, on blogs b_{it}, and z_{it}, controlling for individual and week fixed effects (t-statistics in parenthesis):

$$f_{it} = \alpha_i + \alpha_t + \underset{(13.6)}{0.137} b_{it} + \underset{(0.55)}{0.013} w_{it} - \underset{(-1.43)}{0.067} s_{it} + \underset{(1.34)}{0.030} m_{it}$$

We find that 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 assumption.

3 Empirical Framework

We now discuss the empirical framework we adopt for the estimation of the model. We outline three different approaches, each entailing different assumptions or specifications, and show our results are robust across each of these. First, we outline a linear simontaneous

Variable	Obs	Mean	Std. Dev.	Min	Max
w	11776	2.32	0.44	1.5	3.57
s	11776	0.58	0.24	0	1.67
m	11776	3.17	0.71	1.5	6.00

Table 2: Descriptive Statistics of Wind Variables

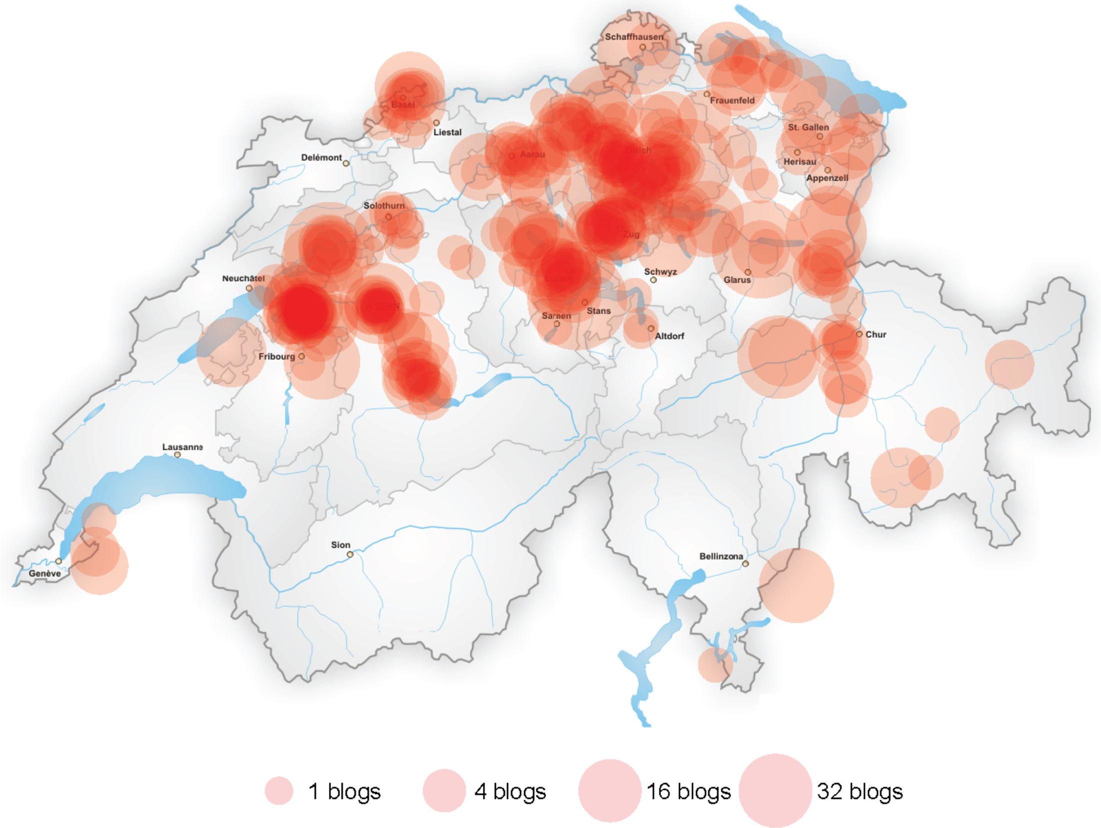

Figure 3: Geographic Distribution of Blogging

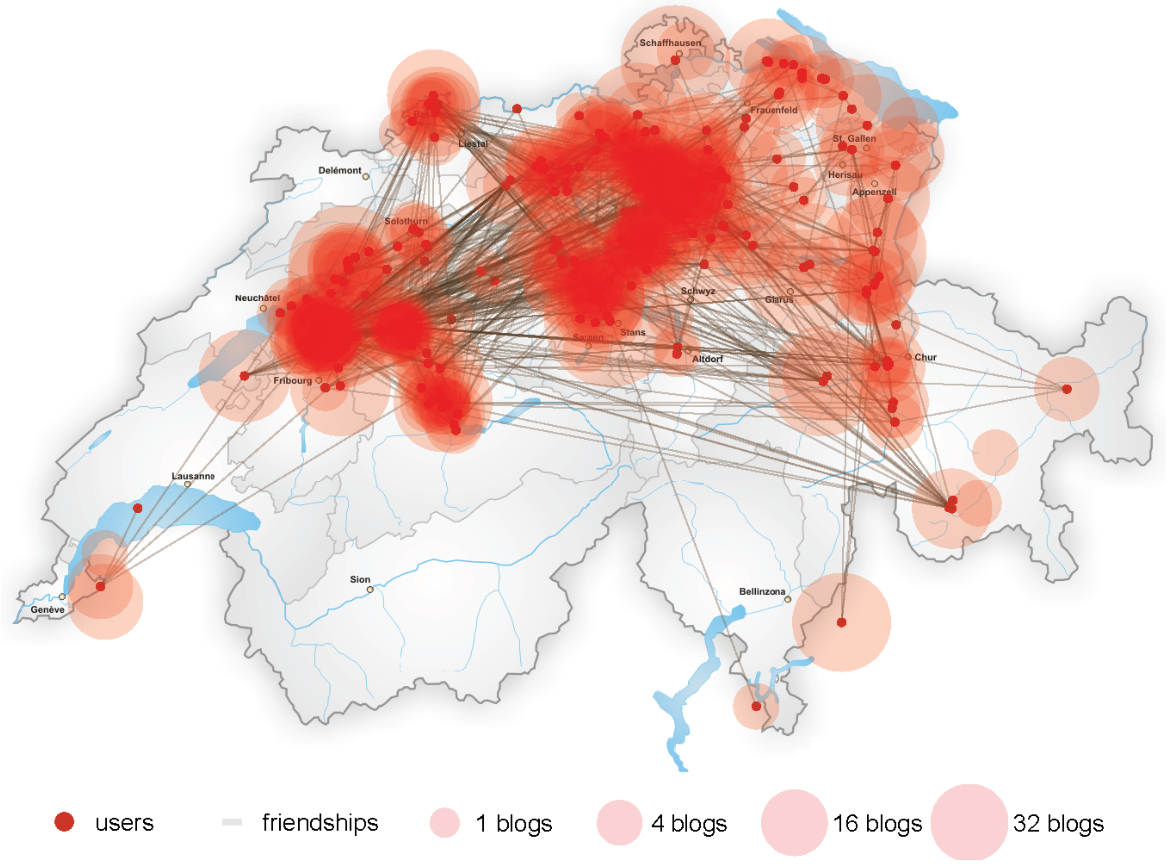

Figure 4: Blogs and Network Ties

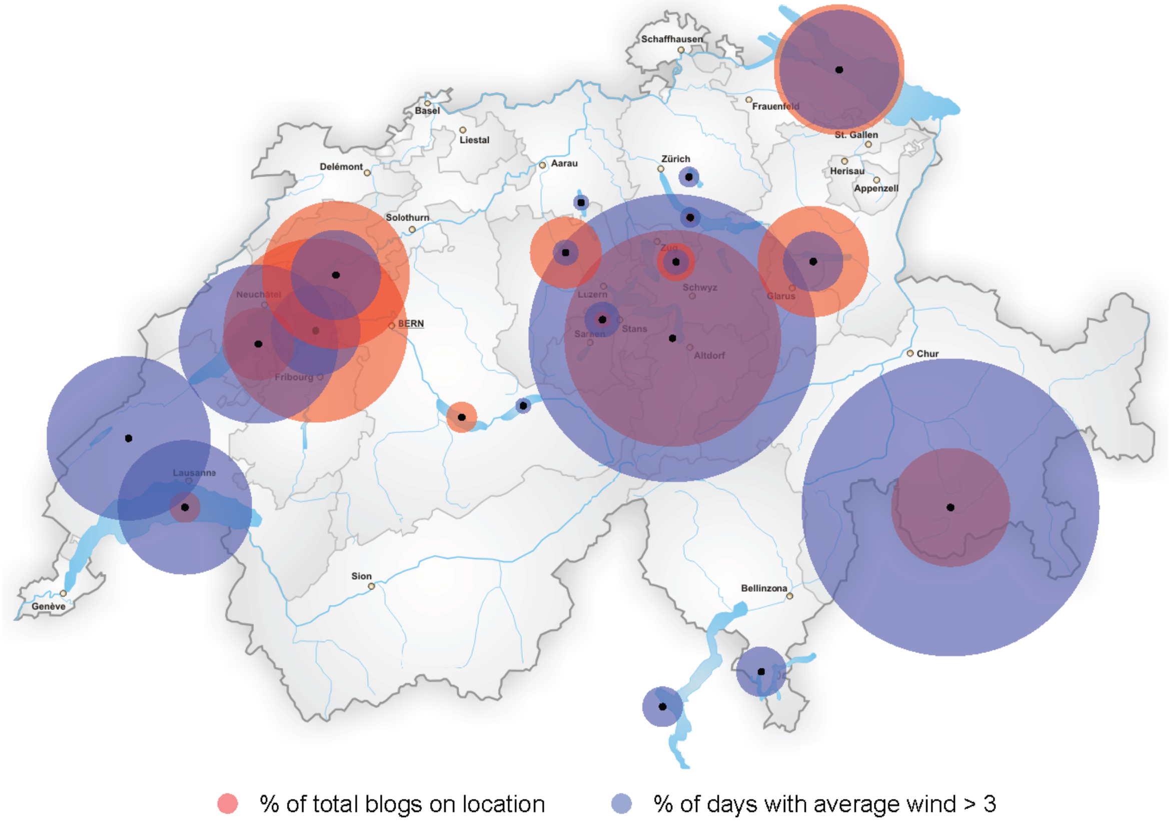

Figure 5: Blogging is Related to Wind-Speeds

equations model of content generation and friendship formation. We discuss how we can estimate this model using two approaches, (a) a control function approach, and (b) a GMM approach. The GMM approach requires access to additional exogenous shifters of *friends*, in addition to the "wind" instruments for *blogs*. We discuss a time-series identification strategy using an Arellano-Bond type estimator that employs lagged *friends* as instruments. We demonstrate that you data satisfies the conditions required for the validity of this instrument. In our second approach, we outline a nonlinear count 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 GMM approach requires using lagged *friends* as instruments. Finally, we show robustness to the lagged instrumental variable. We present a full information maximum likelihood (FIML) approach that does not require access to instruments for the *friends* variable. The FIML model is complicated by the simontaneous nature of the system, and the fact that both dependent variables are count variables. The results section of this draft presents results for the first two approaches. We are currently working on incorporating the results from the FIML approach.

3.1 Linear Model

We start with a linear simontaneous 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 $z_{it} = (w_{it}, s_{it}, m_{it})$, denote the wind measures that affect blogs but is excluded from friends. The linear model is,

$$b_{it} = \alpha_{1i} + \gamma_{1t} + \delta z_{it} + \theta_1 f_{it} + \varepsilon_{1it} \tag{1}$$

$$f_{it} = \alpha_{2i} + \gamma_{2t} + \theta_2 b_{it} + \varepsilon_{2it} \tag{2}$$

The causal effects of interest are $\theta = (\theta_1, \theta_2)$, the marginal effects of blogs and friends 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 arises from homophily or endogenous group formation (e.g. Nair et. al. 2006).[2] 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. Even after controlling for these, equation by equation estimation of this system by OLS is still inconsistent for θ. To see this, consider Equation (2) for f. The RHS variable, b_{it}, is correlated with ε_{2it} because

[2] Endogenous 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 by including a rich specification for heterogeneity (e.g. Hartmann, 2008). Both fixed and random effects serve the role of picking up common aspects of group tastes.

(1), $b_{it} = b_{it}(\varepsilon_{2it})$ from Equation (1), and we do not *a priori* assume that $corr(\varepsilon_{1it}, \varepsilon_{2it}) \neq 0$, and (2) because b_{it} is directly a function of f_{it} by Equation (1). However, as long as $\delta \neq 0$, an exclusion restriction exists, and we can estimate the two parameters by a two step procedure, which we call the control function approach.

3.1.1 Control Function Approach

1. In the first step, estimate Equation (2) via two-stage least squares using z_{it} as instruments for b_{it}. This gives consistent estimates of $\left(\hat{\alpha}_{2i}, \hat{\gamma}_{2t}, \hat{\theta}_2\right)$, as well as $\hat{\varepsilon}_{2it} = f_{it} - \left(\hat{\alpha}_{2i} + \hat{\gamma}_{2t} + \hat{\theta}_2 b_{it}\right)$.

2. In the second step, note that, we can substitute Equation (2) into Equation (1), to obtain,

$$b_{it} = \alpha_{1i} + \gamma_{1t} + \delta z_{it} + \theta_1 \times \left[\hat{\alpha}_{2i} + \hat{\gamma}_{2t} + \hat{\theta}_2 b_{it} + \hat{\varepsilon}_{2it}\right] + \varepsilon_{1it} \tag{3a}$$

$$b_{it} = \underbrace{\frac{\alpha_{1i}}{1-\theta_1\hat{\theta}_2}}_{\tilde{\alpha}_{1i}} + \underbrace{\frac{\gamma_{1t}}{1-\theta_1\hat{\theta}_2}}_{\tilde{\gamma}_{1t}} + \underbrace{\frac{\delta}{1-\theta_1\hat{\theta}_2}}_{\tilde{\delta}} z_{it} + \underbrace{\frac{\theta_1}{1-\theta_1\hat{\theta}_2}}_{\tilde{\theta}_1} \times \left[\hat{\alpha}_{2i} + \hat{\gamma}_{2t} + \hat{\varepsilon}_{2it}\right] + \tilde{\varepsilon}_{1it} \tag{3b}$$

We can now estimate the parameters $\left(\tilde{\alpha}_{1i}, \tilde{\gamma}_{1t}, \tilde{\delta}, \tilde{\theta}_1\right)$ by OLS. Essentially, we have removed the correlation of the endogenous variables with the error term by plugging in the model form for f into the equation. Further, the inclusion of consistent estimates of the residuals $\hat{\varepsilon}_{2it}$ served as a control function for residual correlation with $\tilde{\varepsilon}_{1it}$. Given estimates $\left(\tilde{\alpha}_{1i}, \tilde{\gamma}_{1t}, \tilde{\delta}, \tilde{\theta}_1\right)$, we can recover the primitive parameters as,

$$\alpha_{1i} = \tfrac{\tilde{\alpha}_{1i}}{1+\tilde{\theta}_1\hat{\theta}_2} \quad \gamma_{1i} = \tfrac{\tilde{\gamma}_{1i}}{1+\tilde{\theta}_1\hat{\theta}_2} \quad \delta_{1i} = \tfrac{\tilde{\delta}}{1+\tilde{\theta}_1\hat{\theta}_2} \quad \theta_1 = \tfrac{\tilde{\theta}_1}{1+\tilde{\theta}_1\hat{\theta}_2} \tag{4}$$

Further, standard errors can be obtained in a straightforward way via bootstrapping.

3.1.2 GMM Approach

The GMM approach requires access to additional instruments for f_{it} in Equation (1). In the absence of other exogenous variation, we use a time series identification strategy, and use the number of friends attained in the previous week and the week prior, as instruments for the current number of friends acquired. Thus, the instruments for f_{it} are $h_{it} = (f_{i,t-1}, f_{i,t-2})$. The validity of this identification strategy relies on the fact that $\mathbb{E}(h_{it}, \varepsilon_{1it}) = 0$, i.e. that past tie formation is uncorrelated with current unobservables driving blogging. This will be the case if ε_{1it} is not serially correlated. We present tests showing this is not the case. This is to be expected, as the inclusion of time and user fixed effects picks up much of the source of persistence in the unobservables. With these instruments, we can now base estimation on the moment conditions,

$$\begin{aligned}\mathcal{M}_1 &= \mathbb{E}\left[b_{it} - (\alpha_{1i} + \gamma_{1t} + \delta z_{it} + \theta_1 f_{it}), (z_{it}, h_{it})\right] = 0 && (5)\\ \mathcal{M}_2 &= \mathbb{E}\left[f_{it} - (\alpha_{2i} + \gamma_{2t} + \theta_2 b_{it}), z_{it}\right] = 0 && (6)\end{aligned}$$

The parameters that jointly satisfy these moment conditions minimize the GMM objective function, $\left[\begin{array}{cc}\mathcal{M}_1 & \mathcal{M}_2\end{array}\right]' \mathbb{W} \left[\begin{array}{cc}\mathcal{M}_1 & \mathcal{M}_2\end{array}\right]$, 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.

3.1.3 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 percentage, Δ. 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 $(\Delta\alpha_{2i}) \times \theta_1$, and hence, the return on investment on the intervention, in terms of user-generated content is,[3]

$$ROI = \frac{(\Delta\alpha_{2i})}{c} \times \theta_1 \qquad (7)$$

Now consider the case where we incorporate the feedback from friend formation back into blogs. A small change in f in Equation (2) 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 read off the reduced form of the system in Equation (3b). The incremental user-generated content produced by the intervention is $(\Delta\alpha_{2i}) \times \frac{\theta_1}{1-\theta_1\hat{\theta}_2}$, and hence, the ROI on the intervention is,

$$ROI_{\text{Multiplier}} = \frac{(\Delta\alpha_{2i})}{c} \times \frac{\theta_1}{1-\theta_1\theta_2} \qquad (8)$$

This will be greater than θ_1 as long as both effects are positive ($\theta_1 > 0, \theta_2 > 0$) and $\theta_1\theta_2 < 1$.[4] Moreover, the larger the effect of blogging on link-fromation (i.e., the larger the value of θ_2),

[3] In 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) × (Page impressions per blog).

[4] We find both conditions are true in our empirical analysis.

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 θ_1, θ_2 correctly is key to ROI measurement. Further, the reader should note that spurious correlation between *blogs* and friends, i.e. that fact that $corr\left(\varepsilon_{1it}, \varepsilon_{2it}\right) \neq 0$, does not imply any such multiplier on marketing effort. Only causal effects do.

3.2 A Nonlinear Count Model

We now discuss a nonlinear approach designed to accommodate the count nature of the blogs and friends variables. The nonlinearity of the count models prevents a straightforward way to obtain the reduced form of the model, thereby precluding the "control function" approach outlines above for the linear model. Hence, we pursue estimation of the count model using GMM. We briefly present the GMM count model without individual fixed effects below, and then discuss GMM estimation of the count model with individual fixed effects in further detail.

No user fixed effects We base GMM estimation of the parameters of the count 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 moments. This implies the GMM estimator does not impose the equiproportion property of the Poisson model. Estimation is based on the following moment conditions,

$$\mathcal{M}_1 = \mathbb{E}\left[b_{it} - \exp\left(\alpha_1 + \gamma_{1t} + \delta z_{it} + \theta_1 f_{it}\right), (z_{it}, h_{it})\right] = 0 \quad (9)$$

$$\mathcal{M}_2 = \mathbb{E}\left[f_{it} - \exp\left(\alpha_2 + \gamma_{2t} + \theta_2 b_{it}\right), z_{it}\right] = 0 \quad (10)$$

where, the α-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 count 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,

$$\Pr[f_{it}] = \tfrac{\exp(-\lambda_{2it})\lambda_{2it}^{j}}{f_{it}!} \quad \text{and} \quad \lambda_{2it} = \exp\left(\alpha_{2i} + \gamma_{2t} + \theta_2 b_{it}\right)$$

Consider the log-likelihood of f_{it},

$$\mathcal{L}(\boldsymbol{\alpha}_2, \boldsymbol{\gamma}_2, \theta_2) = \sum_i \sum_t \{-\exp(\alpha_{2i} + \gamma_{2t} + \theta_2 b_{it}) + f_{it} \times (\alpha_{2i} + \gamma_{2t} + \theta_2 b_{it}) - \ln(f_{it}!)\}$$

Setting the first order conditions of $\mathcal{L}(\boldsymbol{\alpha}_2, \boldsymbol{\gamma}_2, \theta_2)$ with respect to α_{2i} equal to 0 implies that,

$$\sum_t \{-\exp(\alpha_{2i} + \gamma_{2t} + \theta_2 b_{it}) + f_{it}\} = 0$$

which implies $\exp(\alpha_{2i}) = \frac{\sum_t f_{it}}{\sum_t \exp(\gamma_{2t} + \theta_2 b_{it})} = \frac{\bar{f}_i}{\bar{\lambda}_{2i}}$ 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 approximate the individual fixed effects (Blundell et. al. 2002). Using this approximation, we now set up estimation of the parameters on the basis of the following moment conditions,

$$\mathcal{M}_1 = \mathbb{E}\left[b_{it} - \frac{\bar{b}_i}{\bar{\lambda}_{1i}} \exp\left(\gamma_{1t} + \delta z_{it} + \theta_1 f_{it}\right), (z_{it}, h_{it})\right] = 0 \tag{11}$$

$$\mathcal{M}_2 = \mathbb{E}\left[f_{it} - \frac{\bar{f}_i}{\bar{\lambda}_{2i}} \exp\left(\gamma_{2t} + \theta_2 b_{it}\right), z_{it}\right] = 0 \tag{12}$$

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

3.3 A Full Information Maximum Likelihood (FIML) Estimator

FIML estimation of the above model is complicated because the system of equations defining friendships and blogging is a nonlinear simontaneous equation system with count-endogenous regressors. To appreciate the complication induced by the count nature of the variables, consider how we would approach the problem if friends f and blogs b were continuous, and were jointly defined by the model,

$$b = g_1\left(f, z, \varepsilon_1; \theta_1\right) \tag{13}$$

$$f = g_2\left(b, \varepsilon_2; \theta_2\right) \tag{14}$$

where g_1, g_2 are nonlinear functions mapping f and b to each other. To find the joint likelihood of (b, f) we would first invert the system to find $(b^* = b^*(z, \varepsilon_1, \varepsilon_2), f^*(z, \varepsilon_1, \varepsilon_2))$ that jointly satisfied Equations (13 and 14), and derive the induced distribution from the stochastic errors $(\varepsilon_1, \varepsilon_2) \longrightarrow (b^*, f^*)$ by change of variable calculus. This transformation is not straightforward when (b, f) are count variables and the relationship between them are defined only in terms of probabilities.

To proceed, we start with the standard assumption of an exponential mean in covariates for the count variables, b_{it} and f_{it},

$$\begin{aligned} \mathbb{E}[b_{it}|f_{it}] &= \exp(\alpha_{1i} + \gamma_{1t} + \delta z_{it} + \theta_1 f_{it}) && (15) \\ \mathbb{E}[f_{it}|b_{it}] &= \exp(\alpha_{2i} + \gamma_{2t} + \theta_2 b_{it}) && (16) \end{aligned}$$

This assumption implies the count variables, b_{it} and f_{it} can be written as a measurement error model,

$$\begin{aligned} b_{it} &= \mathbb{E}[b_{it}|f_{it}] + \varepsilon_{1it} && (17) \\ f_{it} &= \mathbb{E}[f_{it}|b_{it}] + \varepsilon_{2it} && (18) \end{aligned}$$

where, the errors $(\varepsilon_{1it}, \varepsilon_{2it})$ have a joint bivariate *discrete* distribution (e.g., a bivariate Poisson) denoted by $\mathcal{F}_{\varepsilon_1\varepsilon_2}(.,.)$. We can now write the joint likelihood of the data as,

$$\begin{aligned} \Pr(\mathbf{b} = b_{it}, \mathbf{f} = f_{it}) &= \Pr(\mathbb{E}[b_{it}|f_{it}] + \varepsilon_{1it} = b_{it}, \mathbb{E}[f_{it}|b_{it}] + \varepsilon_{2it} = f_{it}) \\ &= \Pr(\varepsilon_{1it} = b_{it} - \exp(\alpha_{1i} + \gamma_{1t} + \delta z_{it} + \theta_1 f_{it}), \varepsilon_{2it} = f_{it} - \exp(\alpha_{2i} + \gamma_{2t} + \theta_2 b_{it})) \\ &= \mathcal{F}_{\varepsilon_1\varepsilon_2}(b_{it} - \exp(\alpha_{1i} + \gamma_{1t} + \delta z_{it} + \theta_1 f_{it}), f_{it} - \exp(\alpha_{2i} + \gamma_{2t} + \theta_2 b_{it})) \end{aligned}$$

One completes the model with a specification for $\mathcal{F}_{\varepsilon_1\varepsilon_2}(.,.)$. A very flexible specification that allows for under or over dispersion, allows for correlation between the two errors, and has a smoothly concave likelihood is the zero-inflated bivariate Poisson (Karlis and Ntzoufras 2005). We are currently working on estimating this model.

3.4 Discussion

This section presented three 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 co-determined. Codetermination complicates the estimation by generating simultaneity. We presented linear and nonlinear versions of the model to handle this co-dependence. Using the exgogenous avriation in wind-speeds, and under the null that lagged friends are valid instruments, the GMM estimator provides consistent estimates of all parameters. As this latter approach involves a pure time-series identification strategy, we also propose a FIML approach that does not require instruments on the friends equation. However, as is standard for FIML, the cost of this approach is that an addiitonal assumption is required on the joint distribution of the errors.

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 control function estimation and the GMM estimation for the linear model. Subsequently, we present results from the nonlinear count model estimated via GMM.

4.1 Linear Model: OLS

Table (3) presents results from OLS regressions of friends on blogs and of blogs on friends. Looking at Table (3), 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 (3), we see that this is indeed the case. Looking at the serial correlation diagnostics, we see that the null of no first order serial correlation is also strongly rejected in the OLS model.

4.2 Linear Model: User fixed effects

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 imposes a very significant stress on the data as the inclusion of both individual and month fixed effects imply that all variation common to individuals

Dependent variable:	f	f	b	b
Procedure:	OLS	OLS	OLS	OLS
b	1.72e-01***	1.68e-01***		
	(2.16e-02)	(2.17e-02)		
f			1.65e-01***	1.61e-01***
			(2.17e-02)	(2.17e-02)
w			1.18e-01***	1.06e-01***
			(2.21e-02)	(2.27e-02)
s			2.93e-01***	2.44e-01***
			(5.09e-02)	(5.02e-02)
m			-6.13e-02**	-4.34e-02
			(2.25e-02)	(2.25e-02)
Constant	7.78e-02***	4.98e-02***	-1.46e-01***	-2.23e-01***
	(3.87e-03)	(6.77e-03)	(2.30e-02)	(2.36e-02)
Month effects	No	Yes	No	Yes
Obs	11776	11776	11776	11776
RMSE	0.450	0.450	0.445	0.443
Serial corr p-val	0.000	0.000	0.000	0.000
$p<0.05$	** $p<0.01$		** $p<0.01$	

Table 3: Linear Model: OLS

within a given month, as well as variation common to months for a given individual, 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 Table (4), we see that this is indeed the case. Looking at the column named "OLS with FE," we see that the effect of blogs on friends has dropped from 0.168 to 0.145 when adding user fixed effects (a 17% decrease). The effect of friends on the other hand has dropped from 0.163 to 0.123 (a 25% decrease). We see that controlling for endogenous group formation 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. In particular, on adding fixed effects, we no longer reject the null of no serial correlation in the residuals. We also see that all variables are significant. For testing statistical significance, note that all tables report robust standard errors reported that have been clustered at the user level.

Dependent variable:	f	f	f	b	b	b
Procedure:	OLS with FE	2SLS	2SLS	OLS with FE	2SLS	2SLS
Model name:		FE	FD		FE	FD
b	1.45e-01***	2.82e-01***	2.43e-01			
	(2.23e-02)	(6.85e-02)	(1.74e-01)			
f				1.23e-01***	1.18e-01	1.02e-01***
				(1.95e-02)	(6.22e-01)	(2.42e-02)
w				1.76e-01***	1.80e-01***	1.54e-01***
				(2.45e-02)	(5.41e-02)	(2.21e-02)
s				2.49e-01***	2.53e-01***	1.53e-01**
				(4.81e-02)	(5.62e-02)	(5.05e-02)
m				-5.76e-02**	-5.26e-02*	-3.83e-02
				(2.11e-02)	(2.51e-02)	(2.18e-02)
Month effects	Yes	Yes	Yes	Yes	Yes	Yes
Individual effects	Yes	Yes	Yes	Yes	Yes	Yes
Obs	11776	11776	11408	11776	11040	11040
RMSE	0.425	0.435	0.607	0.396	0.411	0.540
Weak-id F		64.327	17.485		0.639	880.473
Over-id p-val		0.099	0.306		0.104	0.783
Serial corr p-val	0.460	0.460	0.746	0.034	0.034	0.659
p<0.05						

Table 4: Results from Linear Model: User Fixed Effects

Table (4) also presents estimates of the linear model when instrumenting for friends and blogs using 2SLS applied separately to each of the two equations. We present estimates using both Fixed Effects (FE), and First Differencing (FD) approaches. Recall that the instruments for *blogs* are the wind variables, and the instruments for *friends* are lagged and double-lagged values for the *friends* variable. The latter instruments are valid when the *blogs* are not serially correlated as demonstrated above and in Table (4). 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 2SLS columns in Table (4), we see that this is the case. 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.

With instrumenting, the effect of blogs and friends on each other are significant in either the FE or the FD models. Note that both these estimates are inefficient, as the equations are not estimated jointly (we consider joint estimation via GMM below). We also see that the magnitude of the coefficient on *blogs* in the *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.

4.3 Linear Model: Joint Estimation

Finally, we close the discussion of the linear model by discussing the results from joint estimation of the models. The results are presented in Table (5). Two sets of results are reported, the first corresponding to the "Control Function" approach discussed previously, and the second, corresponding to GMM based on jointly imposing the two moment condi-

Dependent variables:	f/b	f/b
Procedure:	2SLS/Control Function	GMM
b	2.43e-01	2.44e-01*
	(1.53e-01)	(1.16e-01)
f	-8.44e-02	1.07e-01***
	(1.30e-01)	(2.05e-02)
w	1.22e-02	1.47e-02
	(1.89e-02)	(2.27e-02)
s	2.22e-01***	2.18e-01***
	(3.86e-02)	(5.01e-02)
m	-3.46e-02	-3.39e-02
	(1.90e-02)	(2.39e-02)
Month effects	Yes	Yes
Individual effects	Yes (FD)	Yes (FD)
Obs	11776	11040
Over-id p-val		0.677
p<0.05		

Table 5: Results from Linear Model: Joint Estimation

tions in Equations (5 and 6). Standard errors for the control function model are obtained via bootstrapping over individuals with 30 replications. Looking at the table, we see that the control function approach is very inefficient: most of the estimates are not significant. Further, the effect of f on blogs has the wrong sign altogether. Overall, these results suggest the control function approach is not working well for these data. On the other hand, we see that the estimates are significant (and of the expected sign) for the joint GMM model in Table (5). Overall, these result suggest the joint GMM approach delivers more efficient estimates.

4.4 Nonlinear Models

Finally, we present estimates for nonlinear count models. We expect these models to perform better than the linear model as they explicitly take into account the count nature of the data (and hence may produce more efficient estimates). We present the results in Table (6) . As a benchmark, we start by presenting maximum likelihood estimates of Negative Binomial Models of friends and blogs, in which we ignore the endogeneity of either variable. In each, we control for user and month fixed effects (models without fixed effects are not presented for brevity). We see that the effect of *blogs* and *friends* are strongly significant in both models. Third and fifth columns of Table (6) now present results from GMM estimation of the count models using moments based on exponential functions of the means. Looking at these results, we see the basic pattern of the results suggest a feedback effect: blogs have a

Dependent variables:	f	f	b	b
Procedure:	Negative Binomial	GMM	Negative Binomial	GMM
b	4.91e-01***	1.04e+00***		
	(4.41e-02)	(2.22e-01)		
f			2.77e-01***	6.58e-01*
			(3.14e-02)	(2.78e-01)
w			1.02e+00***	9.37e-01***
			(1.50e-01)	(2.43e-01)
s			9.25e-01**	1.57e+00***
			(3.01e-01)	(2.21e-01)
m			-2.75e-02	1.36e+00***
			(1.40e-01)	(2.04e-01)
Month effects	Yes	Yes	Yes	Yes
Individual effects	Yes	Yes	Yes	Yes
Marginal eff. of b				
Marginal eff. of f				
Obs	9472	11776	7648	11040
Over-id p-val		0.596		0.068
p<0.05	** p<0.01	*** p<0.001		

Table 6: Results from Nonlinear Count Model

significant effect on friends and the other way around.

Table (6) also show the marginal effect of the two endogenous variables on each other evaluated at the mean value of these variables in the data. At the sample average, a unit increase in blogs produces 0.08 more friendship links. At the sample average, a unit increase in friends produces 0.04 more blogging content. We plan to use these estimates to compute the implied social multiplier via simulation.

5 Implication for ROI

We now discuss how we use these results for computing the ROI from subsidizing link formation. First, we obtain a rough estimate of the extent to which blogs on a users' webpage drives the number of page-impressions (PI) he obtains. The advertising revenue obtained by Soulrider.com, is roughly 0.002 Swiss Francs (CHF) per PI.[5] Thus, we can translate the effect of an incremental blog onto revenues for the site as (incremental PI generated by additional blogs)× 0.002 CHF. We estimate that (t-statistics in parenthesis):

$$PI_{it} = \alpha_i + \underset{(25.95)}{2.85}\; b_{it}$$

[5] 1 CHF is roughly = 1 USD.

For now, we report the results based on the results from the linear model estimated via GMM (i.e. we use results from the last column of Table (5)). Suppose the website is able to implement a policy that adds R friendship links to the average user. Without considering feedback effects, the incremental effect of adding a friendship link on blogs is $\theta_1 = 0.107$ from Table (5). So the incremental revenue this generates is $2.85 \times R \times 0.107 \times 0.002$ CHF. Following the earlier discussion, we can also compute the effect of feedback in producing a social multiplier for blogging. First, note that when the feedback is incorporated, the full effect of adding a friendship link on blogs is, $\frac{\theta_1}{1-\theta_1\theta_2} = \frac{0.107}{1-0.107\times0.244} = 0.109$, a 3% increase over the effect without the multiplier. Thus, when incorporating the full effect, the incremental revenues generated is $2.85 \times R \times 0.109 \times 0.002$ CHF, which is higher. Essentially, the ROI under the full accounting improves. The full accounting thus enables the website to consider policies for generating content that it would otherwise have considered infeasible relative to the cost.

6 Conclusions

We identify the role of social ties in generating content on online social networks. We 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 form 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 has typically plagued empirical analysis of social interactions. The main challenge in the analysis is to separately identify the causal effect on 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 preliminary results show evidence for significant positive feedback in user generated content. We discuss the implications of the estimates for the management of the content and the growth of the network.

7 References

1. Blundell, R., Rachel Griffith, Frank Windmeijer (2002). "Individual Effects and Dynamics in Count Data Models," Journal of Econometrics (108).

2. Cameron, A.C., and Trivedi, P.K. (1998). "Regression Analysis of Count Data," Cambridge University Press.

3. Chatterjee, P., Donna L. Hoffmann, and Thomas P. Novak (2003). "Modeling the Clickstream: Implications for Web-Based Advertising Efforts," Marketing Science, 22 (4), 520-541.

4. eMarketer (2008a). "Social Networkers Aren't There for Ads," http://www.emarketer.com/Article.aspx?id=1006775

5. eMarketer (2008b). "Social Networks: Millions of Users, Not So Many Marketers," http://www.emarketer.com/Article.aspx?id=1006820

6. Freeman, L. C. (1979). "Centrality in Social Networks: Conceptual clarification," Social Networks 1, 215-239.

7. Friestad, Marian and Peter Wright (1994), 'The persuasion knowledge model: How people cope with persuasion attempts', The Journal of Consumer Research 21(1), 1—31.

8. Glaser, Mark (2006). "Wal-Mart 'Flogs' Par for the PR Course," October 27, PBS News Media, http://www.pbs.org/mediashift/2006/10/wal-mart-flogs-par-for-the-pr-course300.html

9. Godes, David and Dina Mayzlin (2008), "Firm-Created Word-of-Mouth Communication: Evidence from a Field Test," Marketing Science, volume 28, issue 4 (July/August), pp. 721-39.

10. Hansen, Lars Peter, (1982). "Large Sample Properties of Generalized Method of Moments Estimators," Econometrica, vol. 50(4), pages 1029-54, July.

11. Hartmann, Wesley (2008). "Demand Estimation with Social Interactions and the Implications for Targeted Marketing," Marketing Science, *forthcoming.*

12. Hartmann, W., Puneet Manchanda, Harikesh Nair, Matt Bothner, Peter Dodds, Dave Godes, Karthik Hosanagar and Catherine Tucker (2007). "Modeling Social Interactions: Identification, Empirical Methods and Policy Implications," 7th Triennial Choice Symposium Session paper, Marketing Letters, *forthcoming.*

13. Kleibergen, F., and R. Paap. (2006). "Generalized reduced rank tests using the singular value decomposition," Journal of Econometrics 127(1): 97–126.

14. Manski, C. F., (1993). "Identification of Endogenous Social Effects: The Reflection Problem," Review of Economic Studies 60, pp. 531-542.

15. Moffitt, R., (2001). “Policy Interventions, Low-Level Equilibria, and Social Interactions,” in Durlauf, S. and Young, P. (Ed.) Social Dynamics, Brookings Institution Press and MIT Press, 45-82.

16. Oestreicher-Singer, G. and Sundararajan , A. (2008). “The Visible Hand of Social Networks in Electronic Markets,” working paper, New York University, Stern School of Business.

17. Nair, Harikesh, Puneet Manchanda and Tulikaa Bhatia (2006). “Asymmetric Peer Effects in Prescription Behavior: The Role of Opinion Leaders,” Journal of Marketing Research, *forthcoming.*

18. Nielsen (2007). “Nielson Online Reports Topline U.S. Data for November 2007,” http://www.nielsen-online.com/pr/pr_071210.pdf

19. Nielsen (2008a). “Nielsen Online Provides Fastest Growing Social Networks for September 2008,” http://www.nielsen-online.com/pr/pr_081022.pdf

20. Nielsen (2008b). “Nielsen Online Reports Topline U.S. Data for November 2008,” http://www.nielsen-online.com/pr/pr_081216.pdf

21. Trusov, Michael, Randolph E. Bucklin and Koen Pauwels (2009), “Effects of Word-of-Mouth Versus Traditional Marketing: Findings from an Internet Social Networking Site,” Journal of Marketing, Vol. 73 (September), 90-102.

22. Verlegh, Peeter W.J., Celine Verkerk, Mirjam A. Tuk and Ale Smidts (2004). “Customers or sellers? The role of persuasion knowledge in customer referral,” Advances in Consumer Research 31, 304—305.