HOW SOCIAL ARE SOCIAL MOVEMENTS?
SOCIAL TIES, LOCAL NETWORK STRUCTURE, AND CONTINUED PARTICIPATION IN VOLUNTARY ASSOCIATIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Thomas Michael Lento

January 2011

HOW SOCIAL ARE SOCIAL MOVEMENTS?
SOCIAL TIES, LOCAL NETWORK STRUCTURE, AND CONTINUED PARTICIPATION IN VOLUNTARY ASSOCIATIONS

Thomas Michael Lento, Ph.D.

Cornell University 2011

How important are social relationships for contribution to collective action? Existing work on contribution to collective action and participation in social movements has shifted its focus from individual characteristics to social structure as the key to predicting participation, but much of the work on social structure focuses on dyadic interactions as predictors of initial contribution. Building on recent research on patterns of group joining behavior, both online and in academic research settings, this study explores the extent to which attributes of dyadic and local network structure predict continued participation.

The research presented here uses data from Wikipedia, the online encyclopedia, and Wallop, a social networking and personal publishing service, to explore four key questions. First, how do dyadic relationships affect rates of continued participation? Second, what is the relationship between local network structure, such as triadic closure, and subsequent contribution? Third, if triadic closure has an effect, is it the result of structural differences in the network, or is triadic closure encoding dyadic tie strength? Finally, are social network attributes predictive of contribution in task-oriented groups, or are they more important for socially oriented groups?

The results of this study highlight the importance of social relationships with other contributors as predictors of participation in voluntary associations. The

importance of local network structure, which does not appear to be an artifact of tie strength, suggests a possible social affirmation effect where individuals are motivated to contribute by relationships embedded within a social group of other active participants. Furthermore, while social interaction is predictive of increased participation in both socially oriented and task-oriented systems, strong social relationships are negatively correlated with contribution in task-oriented settings. This suggests competing effects, where social relationships with other contributors serve as an important attraction even as they distract the contributor from the task at hand.

BIOGRAPHICAL SKETCH

Thomas M. Lento received his B.A. from Cornell University in the spring of 2000, with a double major in Chemistry and Asian Studies. He received his Ph.D. in Sociology from Cornell University in 2011. He currently resides in California, where he works as a Data Scientist at Facebook, Inc.

For my Grandmother, Frances M. “Marty” Lento, with love and admiration.

ACKNOWLEDGMENTS

This project would not have been possible without the help and support of many people over the years. First and foremost, I would like to thank my advisor, Michael Macy, whose mentorship, advice, and encouragement were critical to the successful completion of this endeavor. The other members of my committee – Douglas Heckathorn, Jon Kleinberg, and David Strang – also provided invaluable insight and inspiration. I would also like to thank my colleagues and collaborators, especially Eric Gleave, Marc Smith, and Howard T. Welser, with whom I worked on precursors to the research presented here, and Beth Hirsh, who provided advice and valuable information on the trade-offs between various statistical approaches. Thanks also to Microsoft Research, and especially Marc Smith, for everything I learned as an intern and a contract researcher, and for the data they provided. Thanks also to Wikipedia and the Wikimedia foundation for making their data freely available for research, and to Gueorgi Kossinets and Dan Cosley for providing a cleanly processed version of the Wikipedia data.

In addition to the advice, mentorship, collaboration, and data support received from those mentioned above, this project also benefited from the direct contribution of several research assistants. Thanks to Linda Tsai-I Chu, Sonny L. Michaud, and Kwame Thomison for their help with initial qualitative analysis and literature searches. Michael Wacker wrote the data processing code for converting raw text data into formatted SQL tables, and Vladimir Barash wrote the initial set of SQL scripts for computing metrics and producing statistical data sets. Without their help the research would have taken far longer than it did – they did a tremendous job, and made it possible for me to focus on other parts of the project while they handled data set construction.

I would like to thank my partner, Eunyun Park, for her love and support throughout the process. She taught me the meaning of greatness, and she eased my cares as I struggled along. My successes are tied to having her in my life, and my greatest achievement is holding her love for as long as I have had it. I would be horribly remiss if I did not thank my family for providing their love and support, advice and encouragement. They are the basis for all that I am, and I can only hope I do them justice. Finally, I would like to express my gratitude to my friends and co-workers for their support and assistance. Thanks in particular to Cameron Marlow, whose clever use of behavioral economics provided the incentive I needed to finish the first draft, and Ana Yang Muller, whose obsession with photo captioning provided me with a weekend work buddy while I made my final revisions.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1:
INTRODUCTION

Overview

How social are social movements? How important are social relationships for contribution to collective action? Although these questions have been considered in the literature, current research on social networks and participation in voluntary associations focuses mainly on the importance of dyadic ties for mobilization. Furthermore, little empirical analysis has been done on the importance of local network structure and contribution to public goods.

This research extends the existing literature on social networks and collective action by providing empirical analysis of the relationship between social network structure and continued participation, rather than focusing strictly on initial contributions. It also examines the importance of elements of local network structure, such as triadic closure, as predictors of participation. This represents a logical continuation of the current direction of research on both social movement mobilization and contribution to collective action. Both of these research areas have progressed towards social structural explanations, but the social network analysis employed in existing studies typically focuses on either dyadic interaction or global network characteristics. By focusing on the local structure surrounding an actor this research seeks to gain additional insight into the connection between social relationships and participation in voluntary associations.

Using social ties to predict participation in a social movement is a relatively recent phenomenon in empirical social movement research, which has moved from individual and task oriented explanations to social structural explanations for mobilization. Early work (Cantril, 1941; Toch, 1965) focuses on personal

characteristics and task affinity, and takes more of a psychological approach to understanding movement participation. Later work (see Snow et al. (1980) for examples) argues that the relationship between the individual and recruiters or recruiting organizations is critical, as participants must be exposed to a movement before they can join it. Finally, starting with Snow et al. (1980)'s work on mobilization and continuing with McAdam (1986)'s study of the Mississippi Freedom Summer, social movement researchers have focused on the importance of dyadic relationships between participants and non-participants, showing that social ties to participants are key predictors of eventual mobilization in high-risk activism.

Collective action research has followed a similar trajectory, starting with Olson (1965)'s work on the value of the outcome to the individual relative to the cost of contribution and progressing through models incorporating global network characteristics (Gould, 1993; Oliver et al., 1988) to models of dyadic relationships and sanctioning (Macy, 1991). Recent models of collective action (Centola & Macy, 2007), along with empirical analysis of group joining behavior (Backstrom et al., 2006), focus on modeling contribution as a form of contagion that spreads across dyadic ties and multiplex interactions. These models suggest that the importance of social ties is more complex than the dyadic influence model currently favored by social movement researchers. Indeed, the power of strong ties as a predictor of participation might actually be an artifact of local network structure, although it is also possible that local network structure is encoding some of the importance of tie strength.

Distinguishing between the effects of tie strength and elements of local network structure, such as triadic closure, is one challenge facing researchers in social movements and collective action today. One of the few empirical studies of local network structure and participation focuses on group formation in two settings, the

online social network LiveJournal and academic research circles (Backstrom et al., 2006). The LiveJournal analysis finds a positive correlation between the number of a user's friends in a group and the user's probability of joining the group at a later time. More importantly, the better connected those friends are to each other, the higher the probability that the user will join, even when holding the number of friends already in the group constant. The analysis of academic citation networks shows similar trends, with dense local structures of co-authorship predicting subsequent publication in a new conference. This suggests that increasing amounts of triadic closure amongst participants is a positive predictor of an individual's likelihood to join a given group.

A second challenge is extending existing work on mobilization to continued participation. Social movement research focuses on mobilization, and theories of collective action focus on the initial contribution to a public good. Even diffusion research focuses on adoption of an innovation without considering the long-term utility of that innovation for adopters (Strang & Macy, 2001). Yet current models of mobilization can be extended to explain future participation of current contributors. McAdam (1986) highlights the fuzziness of the boundaries of a social movement and proposes instead that researchers focus on participation in a particular event. Indeed, the event he chose to study represents an intensification of participation in a larger movement, as the vast majority of participants in the Mississippi Summer program were already active in the Civil Rights movement. Similarly, Gould (1991)'s work on the Paris Commune practically equates mobilization and continued participation. It therefore follows that research on mobilization can be readily extended to explain continued participation and changes in participation in social movements or collective action.

The research presented here seeks to address these challenges by corroborating and extending the findings of the Backstrom et al. (2006). Using data from two online

systems, this study explores the relative importance of dyadic tie strength and local network structure on continued participation. The first system, Wikipedia, is an online encyclopedia produced by contributions from millions of anonymous or semi-anonymous editors. The second, Wallop, is a small social networking and blogging service released by Microsoft Research in 2004. Both of these data sets include detailed social network data and transaction logs recording all contributions to each system.

This research examines several related questions. How do social relationships affect rate of contribution and continued participation? What is the relationship between local structure – particularly triadic closure – and contribution to the system? Is triadic closure important in its own right, or is it simply encoding tie strength? Since Wikipedia and Wallop are quite different this research also explores how the impact of these local network measures differs in different contexts. These effects are compared across a task oriented and socially oriented system – Wikipedia is oriented towards producing an encyclopedia, while Wallop is a social space designed to allow people to interact with one another without a clearly defined mission or group-level goal. Although exploratory in nature, this last question represents an extension of McAdam (1986)'s work comparing high-risk to low-risk activism. There is some evidence that social ties remain important in cases where contribution is lower risk and lower cost (Brown & Brown, 2003). Will the effect vary depending on the inherent level of task orientation present in the movement or voluntary association?

The questions above are explored in detail over the course of the next five chapters. Chapter 2 discusses the opportunities and challenges of using online data for social science research. It highlights the value of using online data for this particular line of research, and sets expectations for the scope of the findings presented here. Chapters 3 and 4 present the research methods and results from analysis of Wikipedia

and Wallop, respectively, while chapter 5 discusses the results and presents opportunities for additional research. The remainder of this chapter draws on literature from social movements, collective action, and diffusion to provide the motivation and expected findings for each of the questions discussed above. Due to the exploratory nature of this research formal hypothesis testing is not present. However, there are clear predictions from the literature and these will be examined in the relevant portions of the analysis.

Social Ties as a Predictor of Continued Participation

Existing research on social relationships and social movement mobilization suggests that an increase in social ties to active movement participants should predict an increase in subsequent participation. In one of the earliest pieces to explore the impact of social ties on mobilization, Snow et al. (1980) find a positive relationship between mobilization and the individual's structural position relative to the center of the movement organization. Pre-existing connections with participants who are actively recruiting for the organization are particularly important predictors of eventual participation.

In his analysis of the Mississippi Freedom Summer project, McAdam (1986) extends Snow et al. (1980)'s work by focusing on the importance of social relationships with other participants in a high-risk activism event, where the cost and inherent risk of contribution is high. His findings show that individuals with strong personal connections to other participants are more likely to join a high-risk social movement activity, while those with strong personal connections to non-participants are less likely to join. This, combined with the fact that weak ties do not have any clear effect on participation, suggests that strong social connections are a key factor in an individual's decision of whether or not to participate in high-risk activism, while

weaker connection to participants are not relevant even if those participants are actively recruiting.

Subsequent work provides further evidence for this effect. McAdam and Paulsen (1993)'s follow-up analysis of the Freedom Summer project highlights the importance of strong ties and pre-existing interpersonal relationships in an individual's decision to participate in high-risk activism. Dixon and Roscigno (2003) find a similar effect on participation in work stoppages and strikes, while Brown and Brown (2003) extend these findings to settings where the risks associated with participation are more moderate.

Studies of mobilization in the face of repression tell a similar story. During the revolution of 1989 participants would only get involved in an uprising against the repressive East German regime if they had the proper incentives. These incentives were embedded in personal relationships, and as a result social ties to other participants were an important part of an individual's decision to participate (Opp & Gern, 1993). More generally, the likelihood of individual mobilization in the face of repression is higher if the individual has connections to a social group that favors protest (Opp & Roehl, 1990).

While these results suggest that social relationships should be key predictors of participation in social movement activities, social groups alone are not sufficient for mobilization. In the Paris Commune of 1871 the pre-existing social structures of the neighborhoods underlying the organization of the Paris National Guard were critical to participation in the rebellion, but without the formal organizational structure of the National Guard organization no uprising would have occurred (Gould, 1991). The same could be said for many mobilization events – without organizational networks the social networks would be unlikely to operate.

Most work on social ties and mobilization focuses on an individual's initial decision to participate in a given social movement activity, assuming that once an individual is part of a movement subsequent contribution is a foregone conclusion. Early theoretical models of network structure and contribution to collective action make similar assumptions (Oliver et al., 1988), but these assumptions may not be valid. Gould (1991)'s research shows that if social and organizational ties to a movement degrade, a participant will eventually cease contributing to the collective action. This suggests that as social relationships change, so too will an individual's level of participation.

While suggestive, Gould's findings are hardly conclusive. Despite assertions that his analysis extends from the initial point of mobilization through some period of sustained participation, the distinction in the case of the Paris Commune is not clear. Indeed, Gould's definition of continued participation is effectively mobilization at the final call to arms, and it is reasonable to claim that his measure of continued participation is actually a measure of the true point of mobilization for the Paris Commune. On the other hand, one could argue that McAdam (1986)'s dependent variable actually measures continuation and intensification of participation in the Freedom Summer project, as he defines the mobilization point as the stage at which participants traveled to Mississippi, not the point at which they first begin attending meetings or sign up as interested participants.

While these indistinct boundaries between mobilization and continued participation highlight McAdam's point that mobilization is a fuzzy concept, the results of the two studies are still instructive in our understanding of how social ties affect continued participation. Although McAdam did not consider network dynamics in his analysis, he does find that ties to other participants within the Mississippi Summer organization are critically important to participation. This is consistent with

Gould's findings around mobilization in the Paris Commune. Based on these findings, and the general body of research that exists around initial mobilization, it is reasonable to expect a relationship between social ties to other participants and an individual's likelihood of participating. More specifically, as an individual establishes social connections with active participants, that individual will tend to increase her own activity in the collective action. Conversely, if an individual develops social ties to fringe members of the organization, or if her friends should reduce their involvement, then she will tend to contribute less or even stop participating altogether.

The Importance of Local Network Structure

Relatively little of the empirical work on mobilization considers the structure of the social networks surrounding the recruits, relying instead on raw counts of the numbers of ties an individual has to others within the organization. Two notable exceptions are Fernandez and McAdam (1988) and Gould (1991). Fernandez and McAdam re-visit the Freedom Summer data and consider network centrality measures as part of their model. They discover an interaction between centrality and recruitment context, where centrality is more important in contexts where existing organizational structures are lacking. Gould examines the structure of relationships between both the neighborhoods and the organizational units in which the mobilized members of the Paris Commune were embedded. His results are striking, and highlight the importance of considering the social and organizational context in which individuals are operating. Although these approaches make excellent use of the available data, they ignore the potential impact of structural factors that can only be observed within the structure of the network immediately surrounding the focal actor.

Research on diffusion suggests that this local network structure can be a more important predictor of individual behavior than dyadic relationships. In their classic

analysis of the diffusion of a medical innovation, Coleman et al. (1957) attribute the spread of the innovation to contagion and interpersonal influence, but later analysis of the same data (Burt, 1987; Strang & Tuma, 1993) suggests that the mechanism for the spread of the innovation through the network was some combination of structural equivalence and social cohesion. While these later findings highlight the importance of an individual's position in a social network, their reliance on structural equivalence may be too restrictive. Other elements of local network structure might be important, and could exert similar influence over individual behavior even though two actors occupy structurally different positions in the network.

Triadic closure in the local network could be sufficient to promote adoption, as it produces the social reinforcement necessary to encourage an otherwise reluctant actor to adopt (Centola & Macy, 2007). Although earlier research shows how densely interconnected local network structures can serve as important mechanisms for social control and norm enforcement (Coleman, 1988), only a handful of studies (Backstrom et al., 2006; Centola & Macy, 2007) directly assess how local network density and triadic closure are related to participation in a voluntary association. Indeed, much of the recent work (Lai & Wong, 2002; Leskovec et al., 2007; Valente, 1996; Watts, 2002; Watts, 1999; Watts, 2004; Watts & Strogatz, 1998) on network structure and adoption or participation focuses on the small world effect, or how connectivity between clusters facilitates diffusion, rather than on the importance of local network clustering for participation. Although the social movement literature is mostly silent on this topic, by drawing on theoretical models of diffusion and collective action it is possible to make predictions about the impact of triadic closure on contribution to a collective action.

Networks and Collective Action

Although most models of collective action do not directly assess triadic closure or other such measures, they do suggest that local network structure is important to the success of a public good. Collective action research has progressed from work focusing on atomic rational actors to interdependent collections of individuals. More recent models focus more closely on local network structure, and suggest a particular path towards gaining a theoretical understanding of how triadic closure might influence contribution.

In one of the earliest pieces of research to consider interdependent actors, Oliver et al. (1985) propose a model in which an individual's calculation of the costs and benefits of contributing changes as others contribute to the public good. By considering differences in resource allocation and heterogeneous interest in producing the collective good, the model highlighted conditions under which contribution was the rational decision. Later studies (see Oliver and Marwell (2001), Oliver (1993) for examples) consider additional structural factors, such as group size (Oliver & Marwell, 1988) and the structure of social (Oliver et al., 1988; Oliver & Myers, 2003) and mass media (Myers, 2000) networks.

Social network models of collective action primarily focus on dyadic influence and global structural factors such as network density and the distribution of social ties (Gould, 1993; Kim & Bearman, 1997; Macy, 1991; Oliver et al., 1988). The earliest of these studies suggests that mobilization is more likely to occur if organizers are centrally located, and denser networks are more likely to mobilize (Oliver et al., 1988). A learning model of collective action that does not include the concept of a core set of organizers (Macy, 1991) produces the opposite result, finding that high levels of density result in uncooperative cliques that inhibit the chain reactions of cooperation necessary to achieve a critical mass. Gould (1993)'s findings suggest

some interdependence between density and centrality, while Kim and Bearman (1997) show that the effect of density on mobilization is dependent on both centrality and resource allocation.

While these studies all highlight the potential importance of network structure as a means of disseminating information about the mobilization effort, most of them do not directly address the role that local structure might play in an individual's decision to contribute to collective action. One notable exception is Macy (1991)'s stochastic learning model, which considers the relative importance of short and long ties, where long ties connect otherwise socially distant actors and short ties connect actors who are part of the same densely interconnected local cluster. This model suggests that long ties connecting individuals from distinct clusters can promote cascades of contribution, but despite an examination of relative density Macy stops just short of directly assessing the importance of structure within the local neighborhoods of the contributors. Chwe (1999) also presents a model that highlights the importance of connections within an individual's local neighborhood as predictors of contribution, but his model is both constrained by an assumption that individuals act in concert and by his conflation of local structure with tie strength. Although it is important as an indicator that local structure matters in assessing mobilization, Chwe's threshold model is difficult to extend. On the other hand, Macy's work is an example of a more general class of social cascade models that can be readily extended to generate predictions about the impact of local structure on contribution to collective action.

Cascade Models

Cascade models capture the spread of some object through a network of agents. The simplest variations, which are based on models of infectious disease

(Anderson & May, 1992; Newman, 2002), posit some probability that an infection will be transmitted from one individual to the next if those individuals are connected in the network. Variations on simple contagion models are frequently used to model the diffusion of information (Adar & Adamic, 2005; Leskovec et al., 2007) and participation in riots and social movements (Feinberg & Johnson, 1988; Johnson & Feinberg, 1977). Theoretical models of social cascades (Centola & Macy, 2007; Watts, 2002) show that network structure is an important factor in the probability of a large cascade, and that under certain conditions increasing triadic closure may allow cascades to happen where they would otherwise be impossible.

The rate at which a contagion will spread is dependent on both the infectiousness of the contagion and the structure of the underlying network. A piece of information will spread rapidly through a densely connected social group just as influenza will spread rapidly through a densely populated neighborhood. In such cases, the local density is the critical factor in the rapid spread of the contagion. A rapid global cascade, where the entire network becomes infected within a short period of time, will obviously happen if the network is densely connected.

Such cascades can also happen even on sparse networks with densely clustered local neighborhoods. In such cases, the infection spreads along bridges that connect otherwise disconnected clusters. Once the infection spreads to a new cluster, the densely connected local neighborhood rapidly transmits it to the other members of the neighborhood. This local infectiousness is important at the neighborhood level, but at a global level the bridge ties are the critical factor in rapidly disseminating the contagion through the network (Centola & Macy, 2007; Watts, 2002).

This echoes Granovetter (1973)’s insight on the strength of weak ties as conduits of information. Strong ties tend to connect individuals to those who are close and connected to others in the same dense cluster of relationships, while weak ties

tend to connect individuals to those who are distant and likely belong to a separate cluster. Of course, the real strength of weak ties is their range – they connect individuals who would otherwise be connected by far longer paths (Centola & Macy, 2007). These bridges provide shortcuts from one cluster to another distant set of nodes, allowing a contagion or a piece of information to spread from one part of the graph to the other.

The importance of long ties also extends to situations where the costs of participation or resistance to an infection are unevenly distributed. Such cases can often be captured with a threshold model (Granovetter, 1978), which assumes that individuals all have some threshold for participation related to the behavior of others in the network. For example, if an individual is choosing whether or not to contribute to a public good, she might base that decision on how many others are contributing. Some threshold models assume that this threshold is based on global participation, and individuals make decisions with perfect information about the activity of others in the network (Granovetter, 1978). Others assume that individuals base their activity on the behaviors of those around them (Chwe, 1999; Macy, 1991; Watts, 2002).

Regardless of the assumptions about how individuals make their decision, threshold models also produce global cascades that spread across bridge ties connecting otherwise disjoint clusters (Macy, 1991; Watts, 2002; Watts, 1999). However, the probability of a global cascade is related to the structure of the network and the distribution of thresholds amongst the nodes. Cascades are less likely in cases where the degree distributions are skewed, but they are more likely in networks with more heterogeneous thresholds. The latter effect is partly due to the higher probability that a low-threshold actor will encounter the contagion and spread it to a new cluster of otherwise unexposed individuals (Watts, 2002).

While these threshold models suggest that local structure – and particularly the range of an individual's ties and the connectedness of his local neighborhood – is important for the spread of a contagion, they do not account for activities that require social support or reinforcement. Activities that carry a high cost, such as the high-risk forms of activism studied by McAdam (1986) may require a larger social support structure before an individual will agree to participate. In these situations hearing about the activity from a single source may not be sufficient to convince an individual to participate. Most threshold models are unable to capture this case, as they assume from the beginning that a certain proportion of the network will agree to participate after just a single exposure. Indeed, it is these vulnerable nodes that allow global cascades to occur in networks populated with relatively high threshold actors.

Centola and Macy (2007) address this gap in the literature with a model of complex contagions, or contagions requiring multiple sources of exposure to a new activity before the focal actor will agree to participate. In this model, the contagion cannot be transmitted between two clusters connected by a single link. The long bridge tie, which is so powerful for transmitting information, becomes the weak link in the chain and prevents the contagion from leaving its source cluster. In order for the cascade to occur, wide bridges between clusters must exist. These wide bridges, which occur when nodes in a new cluster are connected to multiple members of other clusters, allow the contagion to spread through the entire network.

These results suggest that different local structures facilitate the spread of different types of contagion. Simple contagions with low thresholds spread most effectively on networks with many single bridges connecting otherwise disparate clusters, while complex contagions with high thresholds will only spread if there is some degree of triadic closure within and between local clusters. Therefore, in cases where the contagion itself requires a certain degree of social support or social

reinforcement, increased levels of triadic closure should generally facilitate participation.

Empirical analysis highlights the importance of triadic closure as a predictor of an individual's decision to participate in a collective action. In particular, a study of group joining behavior (Backstrom et al., 2006) shows that dyadic ties to group members and triadic closure in an individual's local network are both correlated with the individual's likelihood of joining the group, a finding that supports Centola and Macy (2007)'s core insight. In this study, Backstrom et al. (2006) analyze data from two distinct settings, the online community LiveJournal and the DBLP database of academic publications.

Backstrom et al. (2006) make use of the fact that both of these data sets have include detailed information about an underlying social network and a set of groups available for members of the social network to join. In the LiveJournal data, the network is defined as users and explicit connections formed when one user designates other members of LiveJournal as friends. In the DBLP data nodes in the network represent authors and edges represent co-authorship of a conference paper. The groups in LiveJournal are specific entities which exist within the service, usually organized around topics of interest. Users may join one or more groups in order to interact with other users, and doing so requires an explicit action on the part of the user. In the co-authorship network the groups are conferences, and joining a group is defined as publishing a paper at that conference for the first time. For the sake of simplicity, Backstrom et al. (2006) refer to co-authors as friends and conferences as groups in order to unify the terminology across the two data sets. This convention is also used here.

Backstrom et al. (2006) divide the social network and group joining data derived from LiveJournal and the DBLP into two time periods. They use data on

group membership and social relationships from the first time period to predict the likelihood of an individual joining a group in the second time period. The strongest predictors in the model relate to the number of friends who joined the group prior to the individual in question. The more friends a user has in a given group at time one the more likely the user is to join at time two, although the impact of each incremental friend diminishes after the first few friends have joined the group in question. Furthermore, when the number of friends in the group is held constant, an increase in connections between friends who are members of a given group make a user more likely to join that group. This suggests that some kind of social reinforcement is affecting an individual's decision to join a group, either due to some increase in trust or social coordination.

Triadic Closure, Tie Strength and Individual Participation

The LiveJournal study suggests that triadic closure is an important factor, but there is still the possibility that the underlying mechanism is tie strength rather than local structure. In the theoretical literature closed triads are cited as an indicator of strong dyadic relationships (Granovetter, 1973), a property often used in models to either generate different local network structures (Chwe, 1999; Skvoretz & Fararo, 1996) or to allow researchers to use structural overlap as a measure of tie strength. Indeed, McAdam (1986) used triadic closure as his sole measure for tie strength in his analysis of the Freedom Summer project. While it is entirely possible that McAdam's key finding – that strong ties are predictive of participation – indicates that triadic closure is predictive of participation, it is equally likely that the LiveJournal study is encoding increased influence across one or more strong ties as an effect of triadic closure.

One existing study (Eagle et al., 2010) has examined the impact of triadic closure with controls for tie strength. By analyzing the adoption of a new telephone service among existing British Telecom subscribers, Eagle et al. (2010) are able to replicate Backstrom et al. (2006)'s LiveJournal study. They find a similar relationship between triadic closure and adoption, and triadic closure remains important even after controls for dyadic tie strength are included in the model. Although triadic closure and tie strength are correlated, the persistent importance of local structure in the model suggests that it is an important predictor of adoption in its own right.

The theoretical and empirical research discussed here highlights the importance of local structure as a predictor of individual behavior. A higher rate of triadic closure between prior participants is correlated with an increased probability of an individual joining a group (Backstrom et al., 2006) or adopting a new technology (Eagle et al., 2010). Although existing research focuses on triadic closure and initial adoption or participation, theoretical and empirical work shows closed triads exert more influence than open triads (Bott, 1957; Coleman, 1988) even in persistent interactions. Therefore, as a predictor of continued contribution, triadic closure amongst more active alters should correlate with an increase in participation, while triadic closure amongst less active alters should be predictive of a decline in participation. Furthermore, while tie strength may eliminate some of the effect of triadic closure, prior empirical research (Eagle et al., 2010) suggests that triadic closure will remain important even after controlling for dyadic tie strength.

Empirical Analysis of Task Oriented and Socially Oriented Environments

The research presented here examines the relationship between social ties and local network structure on participation in two different online environments. The first, Wikipedia, is a system where anybody can edit any page on the site, and the goal

of the community is to produce a free online encyclopedia. The second, Wallop, is a social networking and blogging system with no explicit goal. Wikipedia is clearly task oriented, with a highly specific goal and no formal mechanism for social interaction. By contrast, Wallop has no specific goal and is socially oriented with an explicit framework for establishing social connections. Indeed, the primary purpose of Wallop is to share content and interact with other users.

These systems were chosen for this study because they both provide excellent transactional network data. The data sets used here include a complete record of all interactions between users of each system. These interactions span a period of ten months in the Wallop data and over three years in the Wikipedia data, and each interaction is time-stamped. This detailed network data makes it possible to measure changes in tie strength and local network structure over time, a task that is exceedingly difficult – if not impossible – with traditional social network data collection methods (Marsden, 1990; Marsden & Campbell, 1984).

In addition to the detailed social network data provided by these two systems, the fact that one is task oriented while the other is socially oriented provides an interesting area for exploration. McAdam (1986) claims that social ties are most important as predictors of participation in high-risk social activism, and the effects will likely be weaker in low-risk or low-cost settings. As a result, his analysis and most of the studies that follow primarily focus on high-risk activism settings (Dixon & Roscigno, 2003; Fernandez & McAdam, 1988; Gould, 1991; McAdam & Paulsen, 1993; Opp & Gern, 1993; Opp & Roehl, 1990). One of the few studies of social ties and activism in low-risk settings finds social ties are an important predictor of participation in social movement activities organized through churches (Brown & Brown, 2003), which tend to be socially oriented organizations where members seek social support from other members of the church (Becker, 1999) and as such provide a

different mobilization environment from the high-risk activism cases explored by other researchers.

This suggests that in cases of low-risk or low-cost activism social connections will still be important if social relationships are a key feature of the group sponsoring the activity. However, in low or medium risk cases social ties may not have an impact on participation in a group with a strong task orientation and little to no natural emphasis on the importance of social interaction. Since both of the systems studied here feature low risk and low cost contributions, it is possible to examine the relative importance of social relationships outside of the high-risk activism case. Will these social ties be an important predictor of participation? Will they be more important in a setting with a strong social component and no clearly defined good? Intangible goods, such as the social support group that emerges from interaction in an online social network like Wallop, may require a stronger social pull as there is no tangible or permanent outcome. Wikipedia, on the other hand, should have less need for social ties to draw in contributors as it is low cost, low risk, and produces a final outcome of value.

Although Wallop and Wikipedia provide an excellent opportunity to assess the impact of tie strength and local network structure on continued participation in socially oriented and task-oriented systems, there are several critical limitations imposed by the nature of the data itself. First and foremost, there is no reliable demographic data available on the users of these systems. Secondly, while the transactional network data on interaction within each system is as close to flawless as a social network researcher could hope to get, it does not include any interaction that takes place outside of the system. For example, if two Wikipedia users communicate via email, instant messenger, or face-to-face interaction none of that information is captured in the data used for this research. Finally, this research does not include any

form of content analysis or other qualitative approaches that might provide some additional richness and enhance our understanding of the statistical analysis provided here.

Although these limitations are significant, the value of these data are the richly detailed history of interaction provided in each of the two systems. These are large networks of hundreds of thousands or millions of users, and the full history of interaction across the entire network is available for both systems. This provides a tremendous opportunity for research, and makes it possible to empirically test theories that were previously impossible to assess due to the difficulties associated with collecting richly detailed network data. Provided the researchers are careful to keep the limitations of the data in mind when conducting the analysis and interpreting the results, these data sets allow social scientists to make progress on questions that have previously been out of reach.

The next chapter addresses the opportunities and challenges of research on Web data in more detail, beginning with a general discussion of areas where social scientists can derive the greatest benefit from these data sets. It will then address some of the common limitations and considerations that social scientists must keep in mind when using online data before moving into a discussion of the specific strengths and weaknesses of the data for this analysis. Chapter Two concludes with a brief discussion of the approaches taken to mitigate the limitations of the data used in this analysis, and the subsequent chapters provide the detailed methods used on and results derived from both the Wallop and Wikipedia data sets.

CHAPTER 2:

OPPORTUNITES AND CHALLENGES OF USING ONLINE DATA

Online Data and Sociological Research

An increasing amount of social interaction is taking place online through computers and Internet-enabled mobile devices. This interaction is recorded and cataloged in richly detailed transactional data logs, complete with content and time stamps for all activity. The availability of such data represents a unique opportunity to advance the state of the art in research on social interaction and empirical social network analysis.

Despite the promise of these data, sociologists have been slow to use it in their research. This is partly due to the nature of early Internet technology, which was originally a special communication medium available only to a limited, self-selected subset of highly educated and technically skilled individuals. Although services like Usenet and The Well, where Internet users would go in order to socialize with other users from around the world, have been around since the early days of the Internet the popular perception of online activity at that time was that it was not inherently social, and might even be considered anti-social. Therefore, much of the early social science research on Internet use debated whether or not communities and relationships could even be found online (for a review, see Wellman et al. (1996)).

With the rapid spread of the World Wide Web and the rise of social media technologies like online discussion forums, blogs, and wikis in the mid-1990s and early 2000s that perception has largely shifted. Individuals are interacting in various ways through various online media. More importantly for social science researchers, they are recording traces of their interactions and records of their social connections. The widespread popularity of social networking services like MySpace and Facebook,

which began to accumulate tens and hundreds of millions of users between 2005 and 2010, provide large-scale social network data that goes beyond observed interactions to user-provided lists of connections. Some social networking services are as much a map of the user's pre-existing social network as they are a means for individuals to connect with new friends, making this additional layer of network data that much more valuable.

As a result of this new wealth of social network data there are now a number of studies using online data to answer important sociological questions. A study of social networks among political bloggers (Adamic & Glance, 2005) reveals heavily polarized networks, with conservatives generally linked to conservatives and liberals linked to liberals. Work on diffusion through online networks reveals some interesting relationships between the structure of a network and the spread of information (Adar & Adamic, 2005; Leskovec et al., 2006) or the effectiveness of a viral marketing campaign (Leskovec et al., 2007). Additional research on social network dynamics (Adamic & Adar, 2003; Kumar et al., 2003) highlights the mechanisms through which different types of networks form, including how various clusters connect to the main component of a large social network (Kumar et al., 2006).

Other work focuses on processes surrounding group interaction, including the relationship between how individuals are welcomed to a group and their subsequent interactions (Burke et al., 2007; Lampe & Johnston, 2005), how norms are established (Lampe & Johnston, 2005; Postmes et al., 2000), and how individuals come to identify themselves as belonging to a group attempting to provide a public good (Bryant et al., 2005). The relationship between social ties and individual activity is another common area for exploration. Analysis of participation across cultures (Gu et al., 2006; Lento et al., 2006), group joining behavior (Backstrom et al., 2006), and the effects of social

learning on contribution of content (Burke et al., 2009) all suggest that social relationships play a role in individual action.

Work on community formation highlights the value of online data for both qualitative and quantitative methods. Qualitative research on online communities (Herring et al., 2002; Preece & Maloney-Krichmar, 2003) shows the process through which individuals in online groups establish relationships and come to identify themselves as a community. Additional research on conversational trends (Herring et al., 2005) extends this work to highlight how members of a community interact and establish norms for communication and patterns of conversation – patterns which can be important indicators of progress on a social task (Viegas et al., 2007). Work on expert communities combines qualitative and quantitative approaches and highlights the establishment of clearly defined social roles in informal groups (Fisher et al., 2006; Welser et al., 2007). Finally, research combining online data with other data collection approaches examines reasons for participation in online groups (Blood, 2004; Herring et al., 2004; Nardi et al., 2004; Schiano et al., 2004) and assesses social capital and social well-being (Burke et al., 2010; Ellison et al., 2007; Steinfeld et al., 2009; Steinfeld et al., 2008).

Although there is already a wealth of research conducted on online data, the majority of this work has been done by computer and information scientists. In particular, the quantitative research on participation in voluntary associations, diffusion through a network, and polarization has been done by data mining experts who find meaningful patterns in data and develop models based on their empirical observation. Many of their findings are relevant to the field of Sociology, and much of their work would benefit from the input of experts on sociological theory. Yet researchers who start with well-formed sociological theories and test them using online data are exceedingly rare.

This is not to say that sociologists have neglected analysis of online data altogether. Rather, the focus has largely been on either the digital divide or the nature of online social behavior. Digital divide research essentially considers the Internet as a resource, and assesses the extent to which access to this resource is evenly distributed across the population and the impact of unequal access to the Internet (see e.g., Bimber (2000), DiMaggio et al. (2001), Dobransky and Hargittai (2006), Hargittai (2008)). The research on the nature of online behavior primarily considers how computer mediated communication impacts social interaction and assesses the extent to which meaningful social relationships can be formed on the Internet (see e.g., Cummings et al. (2002), Dimmick et al. (2000), Wellman (1996), Wellman et al. (1996)).

While these areas of research are valuable in their own right, they are strictly focused on the Internet as a phenomenon that has an effect on society, not on using online data to assess social science theories. As online data has grown to encompass a larger set of general social activity, sociologists must now shift their focus towards using online interaction as a data source. This richly detailed data on social interaction represents a great opportunity for researchers to answer important theoretical questions that were previously impossible to examine due to difficulties with data collection. Online data are not perfect, and certainly unique challenges arise when a researcher attempts to examine social behavior with data collected from online sources, but it is nevertheless imperative that the discipline begins to embrace this new data and use it as a means of advancing understanding of human social interaction.

The next section highlights the key opportunities of using online data for social science research. The chapter then moves into a discussion of the challenges associated with using this data and some best practices for addressing these

challenges. It concludes with a brief discussion of how these challenges are addressed in the present analysis.

Opportunities for Social Science Research

Perhaps the biggest opportunity online data offers social science research lies in the area of social network analysis. Although social network analysis has provided great insights in the theoretical and empirical literature, it is typically limited by difficulties with data collection. Online data, with its detailed records of social interactions and ready availability of high quality longitudinal network data, can allow social network researchers to conduct analysis that would have been impossible with traditional approaches to data collection. Online data often captures the content of the interactions as well, which provides an excellent resource for qualitative methods in addition to quantitative social network analysis.

Traditional methods for collecting social network data require a great deal of effort on the part of the researcher, and often result in incomplete representations of the ties connecting the individuals involved in the analysis. One prominent method for collecting data involves careful observation of a small, bounded group of individuals in a club or other organization (e.g., Zachary (1977)). Another common approach relies on survey and interview questions, and typically uses a name generator to establish information about the networks surrounding a set of individuals. By comparing the names mentioned by each participant researchers can map out a more complete picture of the full social network (Marsden, 1990). A third approach is to use a form of chain-referral sampling, where the sample begins with a few seed nodes who identify their neighbors in the graph. Each new node then identifies additional nodes, and the graph structure is filled in as new nodes and edges are added to the sample. Respondent-driven sampling (Heckathorn, 1997) is one implementation of this

approach. It has proven useful in mapping out hidden populations, such as injection drug users (Des Jarlais et al., 2007) or HIV patients (Ramirez-Valles et al., 2008), that are difficult to observe or measure directly with traditional survey instruments or other approaches (Heckathorn, 1997; Heckathorn, 2002).

Social scientists have used these methods of network data collection to produce remarkable research on a broad range of topics. Research on social networks and public health issues, ranging from analysis of the connectivity of adolescent sexual networks (Moody, 2002) to patterns of injection drug use (Des Jarlais et al., 2007) to the tendency for obesity to spread across social networks (Christakis & Fowler, 2007), has been conducted using social network data collected via both direct and indirect survey methods. Studies of diffusion have assessed the extent to which structural equivalence, social cohesion, social influence and marketing have affected adoption decisions (Burt, 1987; Coleman et al., 1957; Strang & Tuma, 1993; Van den Bulte & Lilien, 2001). Analysis of recruitment networks for social movements highlights the importance of social ties to other participants (Anheier, 2003; Brown & Brown, 2003; Dixon & Roscigno, 2003; Fernandez & McAdam, 1988; McAdam, 1986; McAdam, 1988; McAdam & Paulsen, 1993; Opp & Gern, 1993) and the intersection of formal and informal networks (Gould, 1991; Hedstrom et al., 2000) as predictors of participation.

Despite the great promise of social network analysis, existing research is often limited by the inherent difficulties associated with collecting data on social relationships. Traditional approaches are labor-intensive and at best provide a limited view of the social network. Observational approaches require a tremendous investment of researcher time, and require intense discipline and meticulous observation. Even with a perfect methodological approach, this method is constrained to only that portion of the network that the researcher is able to observe. This results in network data that

is tightly bounded in both physical and social space. Survey methods are limited by the length of the survey itself, as participants are typically asked to name a few of their friends and acquaintances and answer questions about only those individuals. They are also limited by memory, as people often have difficulty remembering those with whom they interact. Snowball sampling and indirect survey methods might require less direct involvement on the part of the researcher, but they suffer from many of the same limitations as direct survey methods – respondents typically only reveal a small portion of their social network, and edges between non-respondents must be inferred rather than explicitly measured.

The problems with all of these approaches are highlighted when researchers attempt to measure tie strength. The strength of a relationship between two people is exceedingly difficult to measure even with direct observation. Although methods exist for measuring tie strength more effectively (Marsden & Campbell, 1984), these are still less effective than direct observation for cataloging the actual strength of a tie. Furthermore, with so much missing data it is difficult to draw firm conclusions on the relative importance of local structure and tie strength.

Traditional data collection approaches are particularly problematic for research requiring longitudinal analysis of network dynamics. Subsequent surveys or additional observation periods may well yield different networks, but the extent to which this is a result of measurement error as opposed to meaningful change in the network structures is not clear. While statistical methods can provide reasonable estimates of standard errors for several of these approaches (Goel & Salganik, 2010; Heckathorn, 1997; Heckathorn, 2002; Marsden, 1990), access to high-quality data on social interaction would be a tremendous boon to research on questions related diffusion, influence, and other topics where the structure and dynamics of social networks play a prominent role.

Fortunately, data sources derived from online communication provide researchers with access to detailed records of interaction between users, complete with content logs and time stamps for all activity. This level of detail makes it possible to reconstruct the social network that exists within the system. Furthermore, the presence of time stamps on each communication or interaction event allows for dynamic analysis, as the network can be generated for any given point in time.

These detailed records of online social activity allow researchers to examine how an individual's local network changes over time, and how that individual's behavior changes over the same period. Researchers can also use this data to measure the frequency and regularity of interaction between two individuals, which can provide some behavioral measures of tie strength. The resulting data provides an accurate and complete picture of an interaction-based social network – i.e., a network derived from actual communication or social interaction. Such data are useful for analysis of dynamic network effects and the importance of local structure for the diffusion of information, influence, or social control.

Beyond network analysis, online data are also useful for qualitative researchers. The richly detailed history of interaction, complete with transcripts of communication, allows a researcher to observe a social setting without establishing a physical presence. This makes it possible to conduct post-hoc ethnographic analysis with minimal risk of altering the behavior of the subjects. The persistence of the original content also makes it possible to use multiple researchers to examine and code the raw data. By combining the qualitative research opportunities with the available data on social interactions, researchers can effectively study the importance of both tie strength and tie valence on individual behavior.

With access to such detailed data on social interaction, researchers can assess theoretical work that has previously been impossible to study. Studies of contagion

through networks have been done through online sources (Adar & Adamic, 2005; Leskovec et al., 2007). It is also possible to use online data for research on network structure and polarization (Adamic & Glance, 2005), norms (Lampe & Johnston, 2005; Postmes et al., 2000), influence (Kale et al., 2007; Leskovec et al., 2006) and social control (Herring et al., 2002; Kollock & Smith, 1996). Finally, novel experiments (e.g., Goel et al. (2009), Salganik et al. (2006), Salganik and Watts (2009)) can be conducted through online services with the assurance of high-quality data on both the individual behaviors and group interactions found amongst the participants.

Challenges of Using Online Data

Although it provides detailed, time-stamped records of individual actions, including communication and interaction with other users, online data are far from perfect. Findings from online data may not generalize well to other settings, partly because of the nature of the online medium and partly due to a lack of demographic data, a flaw affecting a large proportion of available online data. The scope of the data is also limited, as data on interactions outside of the online space is rarely available. Finally, the technical challenges to effectively analyzing online data are significant.

Limitations of Online Data

One concern about using online data is whether or not research findings based on online social interaction generalize to other settings. This concern rises in part from the impact the online environment has on behavior. Individual behavior and forms of social interaction can differ from one online space to the next, as variations in user interface and degree of anonymity create different modes of communication and standards of usage. Although there is a diverse range of online environments, nearly

all of them differ quite sharply from face-to-face interaction, and individual actions in online settings might not translate to similar offline behavior.

Another issue is sampling bias. Even with access to data about the full population of users of a given online service, researchers using these data must always be mindful of the potential bias in their sample. There is no guarantee that the user base of that service is representative of the online population as a whole, much less of the population of a nation or particular social region. To make matters worse, online data often lacks demographic characteristics, as many services do not require users to provide such information. This makes it difficult to assess sampling bias or examine questions where gender, age, or other demographic factors might have a strong influence on individual behavior.

Although traditional social network data collection methods have their weaknesses, and although online sources tend to provide more complete network data, the traditional methods often do come with some demographic data about each individual in the network. An additional advantage of traditional approaches is their flexibility. Survey methods in particular make it possible for researchers to assess multiple networks, and in particular provide them with data on an individual's social network as seen by the individual. Online data usually provides information on an interaction network within the context of a particular online service. Data on interactions in other contexts is often missing, so even if two users communicate via email or face to face they may appear to be disconnected in a given online setting.

The technical challenges to using online data may be more daunting than the challenges arising from the nature of the data itself. Online data are typically orders of magnitude larger than the quantitative data typically collected and used by social scientists, the level of noise in the data can be comparatively extreme, and the data

often features heavily skewed distributions. These factors can complicate the process of analyzing the data and interpreting the results.

Technical Challenges of Using Online Data

Most of the purely technical challenges of handling online data are a direct result of the volume of information available. Social scientists often work with relatively small numbers of observations. Large surveys, like the General Social Survey, typically receive tens or hundreds of thousands of response. Large network data sets typically consist of a few hundred or a few thousand nodes, and experimental and observational data are smaller still. By contrast, there are nearly ten million registered users at Wikipedia, and over one million nodes in the social network derived for this study. Wallop, a tiny service by Internet standards, attracted over 100,000 registered users, 50,000 of which were active in the social network derived for this study. Larger services, such as Facebook, GMail, MySpace or Yahoo! Groups can have hundreds of millions of users.

Processing such a large amount of data often requires some degree of technical skill. In some cases, data reduction or sampling methods may be sufficient, but for studies where network analysis is of prime importance sampling before processing the network data is often impossible. In these cases, researchers must write scripts to process the data and compute relevant statistics, and special computer hardware is often required to run these scripts over the full data set.

The second problem is data cleaning. Just as survey researchers must deal with missing values or invalid responses, researchers using online data must contend with a certain amount of noise in the data. Various problems might exist in the data, including system failures, persistent bugs, duplicate data, and even small but important annoyances like time stamps that are sensitive to shifts in daylight savings time.

Furthermore, once all of the system-level problems with the data are worked out, there are user-level issues to consider. Some small percentage of users in any given system will be spammers or robots, and while automated algorithms for detecting these users are available (see e.g., Kolari et al. (2006)) it is not always easy to distinguish these from legitimate users who happen to be extreme outliers. Furthermore, even relatively ordinary users are prone to unexpected behavior. For example, certain subsets of the MySpace user population systematically misrepresent their age (Caverlee & Webb, 2008), an action that might affect certain types of analysis. In order to conduct solid, reproducible, trustworthy results researchers using online data sets must account for all of these possible sources of error.

Statistical analysis of online data can also be problematic, in part because normal distributions are quite rare. Much of the large-volume online data available to researchers prominently features power-law, log normal, or other heavy-tailed distributions. Therefore, many of the assumptions of parametric tests are invalid, and researchers must either normalize the data or use non-parametric tests as appropriate in order to minimize the risk of encoding spurious variation in the behavior of users in the tail of the distribution.

Researchers using online data must also take care when interpreting the results. With such a large volume of data marginal effects can be statistically significant. While using p-values as a measure of importance is generally inadvisable, it is particularly important to avoid this trap with large scale data, and the magnitude of the effect must be included in any examination of the importance of a given result. Furthermore, decisions made earlier in the research process can often influence the results, and this must be accounted for in any final interpretation of the data.

Overcoming the Challenges of Working with Online Data

Many of the challenges posed by online data are not new to social science research. Traditional data sources, such as surveys, lab experiments, and participant observation studies, all suffer from a variety of shortcomings. Many of these shortcomings also apply to online data sources, and therefore the established methods for dealing with problems of scope or sampling bias may also be useful with online data.

Overcoming the Limitations of Online Data

Concerns about whether or not results obtained from analyzing online data will generalize to the face-to-face population are well founded, but these same issues exist for all data collection methods. Do experimental results generalize effectively from the participants or lab environment to the general population? Do observational studies generalize beyond the group under observation? Do surveys accurately capture the effects they are attempting to measure?

Sociologists must always be careful to discuss results within an appropriate scope, and to recognize when a finding might be an artifact of the methodology employed. This is especially true when online data are involved, since a phenomenon observed online might not occur in either a face-to-face or different online environment. One way to contend with this problem is to study multiple environments by using data from multiple online systems or, if possible, from a mixture of online and offline settings. If the research results are robust then they can be reported with greater confidence.

Another drawback of online data is the lack of demographic data, but this is not necessarily as serious as it might seem. In most cases, online data provides excellent structural information but little to no information about demographics, while

survey data provides excellent demographic data but little to no information about social structure. Just as it would not be appropriate to use most survey data to answer questions about the importance of social relationships as predictors of individual behavior, it is not appropriate to use online data to answer questions about demographic factors. However, when presented with a question with the appropriate scope, online data can be quite useful.

Furthermore, it is possible to use statistical methods that control for unobserved heterogeneity (Wrigley, 1990; Zohoori & Savitz, 1997). Depending on the missing demographic data and the nature of the study, these models might be sufficient to overcome this lack of information. For cases where such approaches are not appropriate, there are now online data sources that include basic demographic data, which makes it possible to control for demographic effects directly.

One final shortcoming of online data is the limited nature of the network data. Although the interaction network is complete and detailed, there is no additional information about the interactions or latent relationships between the individuals included in the data. While this is a critical issue, similar limitations exist to some extent in all network data, and most existing data collection methods also suffer from problems with missing data. As with concerns about generalizing results from online data analysis to offline contexts, analyzing data from multiple environments and interpreting the data within the proper scope will substantially reduce the impact of this limitation.

Overcoming Technical Challenges

Some of the technical challenges posed by online data require the use of techniques not often employed in social science research. In particular, processing data at the scale of a large online service demands technical skill and access to powerful

hardware. The ability to write efficient scripts to query databases and process raw text logs is necessary, and some facility with distributed data processing is often a great help. By using distributed computing technologies like Hadoop map/reduce it is possible to process terabytes of data in a reasonably short period of time, and the programming requirements are not particularly severe. While such programmatic data processing has typically been the realm of the computer scientist, the recent availability of large-scale data for sociological research means social scientists would do well to develop the technical skills necessary to do at least do basic querying and processing of the source data.

Most online data contains a fair amount of systematic bias, and most of the relevant activity distributions will be heavy-tailed. Cleaning the data and contending with the skewed distributions requires a rigorous approach to understanding the data. Just as a survey researcher might look at descriptive statistics to see how responses are distributed and how the data breaks down across demographic categories, researchers using online data must do descriptive analysis on the full data set. By considering frequency distributions and other large-scale distributions of relevant metrics, it is possible to gain an understanding of how variables are distributed and which summary statistics and normalization techniques are most useful. From there a researcher can look for signs of systematic variance. For example, extreme outliers and perfectly consistent behavior might be signs that a particular set of users is actually a set of spammers or robots.

Furthermore, just as survey researchers must understand the sampling methodology and question phrasing used in the survey that generated the data, researchers using online data must understand the system thoroughly. This makes it easier to see errors and possible sources of bias, and it helps to understand why users might favor a particular feature. A working knowledge of the system that generated

the data will also make it easier to understand the underlying structure of the data, which in turn aids in converting that data to something more directly useful for answering a particular research question.

Finally, analyzing and interpreting the data requires care. With such large data volumes and heavily skewed distributions, online data can be unforgiving. Minor assumptions can cause serious problems in the final analysis, as everything from data processing to variable construction to test selection impacts the final results. Understanding sources of error and how they impact the final analysis is critical to proper interpretation of results from online data. Choosing sound normalization techniques or using non-parametric statistical tests will help reduce possible sources of error. Understanding the difference between meaningful and statistically significant is also critical here. Significance is often achieved for marginal effects due to data volume, so it is important to consider both significance and magnitude when looking at an effect. Finally, keeping the interpretation of the data within the appropriate scope will help reduce the likelihood of misinterpreting the results of an analysis.

Use of Online Data in this Study

Although online data are not appropriate for all analyses, it has much to offer. The detailed log of events and interactions, including associated content and time stamps, makes it possible to conduct network analysis at an unprecedented scale. Furthermore, many of the drawbacks of using online data are not fundamentally different from the limitations of more traditional methods of data collection, such as surveys and participant observation. Problems with scope exist in all research, as do questions about how representative a given sample might be of the larger population. Even in cases where the limitations may be a bit different, they are no more severe than those found in other data sources. Furthermore, social scientists have learned how

to address the problems found in other data sources, and much of that knowledge applies when considering approaches to online data sources. Indeed, a rigorous, thoughtful, statistically and scientifically sound approach to the data should yield solid, reproducible, comprehensible results.

Given the advantages of online data, particularly with respect to network measurement, it is a natural choice for an assessment of the relationship between local network structure, tie strength, and continued participation in a voluntary association. In this study, the challenges of online data are addressed primarily by keeping the research within an appropriate scope. The research questions and expected observations outlined in the first chapter are generated from earlier research in the fields of both social science and computer and information science. This provides a theoretical and methodological grounding for the research.

The primary analysis is based on prior work, and two systems with distinct interfaces are studied. Therefore, any results that remain consistent across these systems are unlikely to be an artifact of interface elements or other system-specific attributes. Furthermore, the data from both systems was explored at length prior to the final analysis, and variables were therefore normalized based on a solid understanding of the underlying distributions. The variables themselves were carefully considered both within the context of the two systems studied here and within the overall context of the research. Whenever possible, variable selection and normalization was conducted the same way in both cases in order to ensure consistency between the analyses, and deviations were only considered when absolutely necessary. Finally, the results were interpreted with the limitations of the data in mind, and the effects of unmeasured demographics, unfiltered robots, or other possible biases were taken into account before the conclusions were drawn. In particular, the lack of demographic data could cause a bias in the interpretation of the network results, as it is conceivable

that triadic closure could be encoding homophily. If demographically similar individuals are more likely to be connected, and connections to demographically similar alters lead an individual to increase production, triadic closure might be an artifact of homophily, and this possibility must be considered in the final analysis.

The technical challenges associated with this research are not particularly onerous. The volume of relevant data is small enough that all relevant information can be stored in a database on a single machine, and many of the queries can be run in memory. Once the raw data was parsed and loaded into the database for more convenient querying, it was a fairly straightforward process to generate metrics and conduct statistical analysis from the underlying data. The only exceptions were certain computationally intense measures, such as triadic closure metrics, which were difficult to compute on the available hardware.

The next two chapters describe this process in more detail. Chapter 3 outlines the Wikipedia analysis, starting with a description of the Wikipedia service and moving into a detailed description of the analysis methodology used before concluding with a presentation and interpretation of the results. Chapter 4 does the same for the Wallop data, with the addition of some minor comparisons between the Wallop and Wikipedia results.

CHAPTER 3:
ANALYSIS OF WIKIPEDIA CONTRIBUTIONS

Wikipedia Analysis

The analysis presented here uses social interaction data to predict how individuals change their levels of contribution to a public good over time. In particular, it focuses on measures of local network structure and tie strength as predictors of individual activity. This approach requires a longitudinal data set with data on contribution rates and detailed information about the social network structure surrounding each individual contributor. Wikipedia, the free online encyclopedia, is one of the best available sources of data for this research.

Data on Wikipedia activity, which is freely available, includes a full log of every editing action taken on the site, with time stamps and user identification codes. This makes it possible to reconstruct both the contribution and social interaction history for each user. Although it is lacking the demographic data necessary to include a complete set of control variables, the granularity of the activity and interaction data makes it well suited for an initial assessment of social interaction and contribution to the public good.

The first step in this analysis, then, is to examine the importance of social interaction within Wikipedia as a predictor of contribution to the encyclopedia portion of the site. The remainder of this chapter discusses a time-series analysis of the relationship between social interaction in the user talk namespace and subsequent changes in frequency of contribution to the main article namespace on Wikipedia. It begins with some relevant details about Wikipedia itself and moves into descriptions of data collection, data processing, and sampling methodologies. Subsequent sections describe methods of extracting network and activity variables from the data, and

explain the statistical models used in the analysis before presenting the final results of the study. The chapter concludes with a brief discussion of the findings.

About Wikipedia

Wikipedia is a free online encyclopedia produced through contributions from thousands of users. One of the core principles of Wikipedia is that anyone can edit any page, and no expertise is required or expected from contributors to the encyclopedia. While some users are blocked due to bad behavior, in principle anyone can edit the site either anonymously or by registering for an account. Even blocked users can re-gain access to the site by creating a new account.

Founded in 2001 by Jimmy Wales, Wikipedia has turned into a major undertaking, and is arguably one of the greatest success stories for the power of user-generated content. The English language version of the site has approximately 3 million articles and over 17 million total pages. Roughly 10 million registered users and an unknown number of anonymous users have made more than 311 million edits to the English language version of the site. These contributions add up to approximately 3 million articles and over 17 million total pages. Although the English version is by far the largest, Wikipedia is available in many other languages. There are 11 other versions of Wikipedia with over 250,000 articles, including German, Spanish, French, Italian, and Japanese. Although some of these numbers are staggering, it is worth noting that the active user population of Wikipedia is quite a bit smaller than the total user base. Of the 10 million registered users, only 155,000 are counted as monthly active users.

At the core of the site is the encyclopedia. Each article in Wikipedia is created through the collective edits of one or more users. Although there are general guidelines in place for language, style, citations, neutrality, and the organization of the

encyclopedia there is relatively little top-down enforcement or oversight. The state of an article at any given point in time – including the content, writing style, and presence or absence of references and sources – is the result of user consensus achieved through an organic collaborative process of writing, editing, and refining the article.

In the trivial case, where one user or a small set of like-minded users is editing a given article, consensus formation is a simple matter of adding text to the article. However, in more complex situations consensus formation often occurs through a separate set of interactions. These interactions take place on the talk page associated with the relevant article. Every article has a talk page associated with it, and the talk page is a place where users can ask questions about the article, explain their reasoning behind a given change, express their opinions about another user's changes, or follow the full history of the discourse behind the development of the article. These article talk pages are, essentially, where much of the work of generating Wikipedia takes place. They are spaces where users can interact directly with one another and discuss the task at hand.

Although each article talk page is linked to a specific article, they are all collected and stored in their own namespace. Namespaces in Wikipedia serve to separate the various parts of the service along some logical division. The Wikipedia service is broken down into several namespaces that roughly correspond to a specific function within the site.

Most of these namespaces serve to organize contributors around certain goals, or to organize content in order to make it easier to find. There is a namespace dedicated to Wikipedia projects, which may be small projects within Wikipedia or a page dedicated to the Wikipedia project itself. Pages in this namespace typically deal with organizational details around a given project, or with policy decisions within

Wikipedia. There are also namespaces dedicated to making it possible to find articles on particular subjects, listing article categories, or providing help files. The file namespace keeps all files attached to articles, such as images, along with some metadata about each file, while the MediaWiki/Template namespace stores tools like page templates or Wikipedia indexes.

One of the biggest namespaces is the user namespace, where the pages associated with each user account are stored. As with all of Wikipedia, anyone can edit these pages, but by convention they are only rarely edited by anyone other than the user owning the page. User pages are often sparsely populated, but when they are used they are similar to a personal profile. Users typically put some small pieces of information about themselves on their user page, ranging from personal information, like favorite books or movies, to professional information, like education or employment data, to Wikipedia-specific information, like the articles the user tends to edit. Some users also place reference materials or to-do lists on their user pages, and others place images or other decorations.

As with articles, each user page has its own talk page associated with it, and these user talk pages are stored in their own namespace. Such pages are typically used for commentary and communication, which may take the form of discussion about the user and his/her interests. In other cases user talk pages can host directed messages posted in public view, as on a personal bulletin board. Messages on user talk pages may be related to Wikipedia, e.g. reprimands for bad behavior, pointers about proper formatting or instructions on finding articles of interest to edit, but many are direct communications about other topics.

These user talk pages provide a social environment for users to interact with one another. They are part of the social space in Wikipedia, the area where Wikipedia editors establish a form of community. This community space is separate from the

encyclopedia itself, but it is still directly connected to the site. Other spaces, such as the article talk namespace, are more directly tied to the Wikipedia project, and support a more task-oriented form of interaction.

Data and Methods

The complete edit history for Wikipedia is publicly available for download from the Wikimedia Foundation, the nonprofit responsible for maintaining and hosting Wikipedia. The Wikimedia Foundation periodically generates a complete dump of Wikipedia's back-end database, including all edits and publicly visible user account data. This database snapshot is then made available for download as an XML file. The research presented in this chapter is based on the Wikipedia data dump from February of 2008.

Once the data was downloaded from the Wikimedia Foundation, the raw XML data was parsed and re-organized into flat text files, with separate files for each namespace. This task was accomplished through a combination of freely available MediaWiki parsers and custom Perl scripts[1]. A Java program[2] then converted these text files to a set of formatted files designed for convenient loading into a SQL database. The database tables where then queried with a set of SQL scripts[3] specifically designed to generate the relevant data sets, which were finally analyzed using a combination of Stata and R scripts.

---

[1] I am grateful to Dan Cosley and Gueorgi Kossinets, who wrote the relevant Perl scripts and did most of the work associated with parsing the raw Wikipedia data dump.

[2] This code was written by Michael Wacker, an undergraduate research assistant.

[3] Vladimir Barash wrote most of the initial SQL queries. I am grateful for his assistance, as it saved me quite a bit of time and trouble. All code is available upon request.

Deriving Social Network Data from Wikipedia

Connections between Wikipedia users can be measured in several ways. The most obvious way is to use a bipartite measure, in which users are linked when they edit the same article. This measure captures a fair number of connections with no meaningful social weight, as users do not typically address each other on the article pages and the other editors of those pages are not immediately visible to a user navigating Wikipedia. Therefore, although this measure suggests some probability of overlapping interests it is largely coincidental and therefore not suitable for the purposes of this study.

Another approach is to focus on task-oriented interaction. Wikipedia users can be linked by interaction on article talk pages. Although such interaction is often coincidental – a user writing on an article talk page may not be communicating directly with any of the other users with content on that page – it is possible to infer direct interaction between two users by filtering for repeated exchanges over relatively short periods of time. These exchanges tend to be task-oriented in nature, as users interacting on article talk pages are almost always doing so because they are discussing the article in question. Similarly, users participating in the same set of projects may interact in the project namespace. Although this may represent a stronger overlap in interests, or even a stronger social bond for users participating in smaller projects, this is still a task-oriented form of interaction and may represent a different set of connections than a measure based on social interaction.

A third approach focuses on direct communication between two users. In Wikipedia, users accomplish this by editing each other's user talk pages. Although some of these interactions are task oriented, many are purely social and indicate some relationship outside of the context of generating the encyclopedia. Furthermore, this measure is more direct than the others, as an edit to another user's talk page is

typically associated with direct communication with that user. While editing another user's talk page does not indicate any particular closeness between the editor and the user, frequent interactions and two-way connections may be indicative of a stronger social connection. This study adopts the third approach, and defines nodes in the Wikipedia user network as registered and unregistered accounts and edges as direct communication through user talk pages.

Edges are defined as direct communication through user talk pages because this study is focused on the impact of social interaction on changes in participation. In particular, this research explores whether or not an individual is affected by the level of participation of his/her friends. Within the context of Wikipedia, only direct communication through user talk pages indicates any relationship outside of the context of producing the encyclopedia. Task-oriented measures like collaboration on articles or interaction on article talk pages will capture the influence of other contributors, whether or not they have any social relationship with the user. Therefore, focusing on interaction in the socially oriented areas of the site is more appropriate for assessing the research questions for this study.

Sampling

The raw Wikipedia data includes every edit made to Wikipedia dating back to the founding of the site in 2001, including edits to all namespaces and edits made by both registered and unregistered accounts. However, some of these data are not useful in the context of this study, and therefore two specific subsets of Wikipedia users are excluded from the data. First, anonymous accounts are not included in the analysis. Although anonymous accounts are identified by an IP address, it is fairly common for many individuals to share a computer with a given IP. This is particularly true for schools, businesses, libraries, and other organizations with many computers sharing a

single Internet connection. Removing anonymous accounts from the user sample helps ensure activity associated with an individual account is tied to an individual person. Second, users who did not send or receive a message within the user talk network are not included in the study. Since the primary goal of this analysis is to examine the relationship between social ties and subsequent changes in participation rates, restricting the sample to users engaging in social interaction on Wikipedia is a reasonable, if somewhat aggressive, restriction.

Much as the users included in the sample are restricted to those meeting certain requirements, the time window used for the analysis is also limited to a subset of the available data. Although the Wikipedia data includes the full history of edits made to both the article user talk pages, the user talk network did not gain popularity amongst Wikipedia users until the middle of 2004. Since the key independent variables in this analysis are derived from interaction on the user talk pages, the analysis only uses data from the middle of 2004 through the end of 2007.[4] In all, roughly 39 months worth of data are included in the sample. This represents approximately half of the active time span represented in the data, but given Wikipedia's growth over time it is substantially more than half of the total available edits.

The final restriction on the data has to do with the namespaces included in the analysis. All user activity is measured based on edits to the article namespace. Therefore, contributions to Wikipedia are defined as contributions to the encyclopedia itself, rather than contributions to supporting namespaces like projects. While it is reasonable to consider participation in projects, the addition of new tools, or even interactions on an article talk page as meaningful contributions to the Wikipedia project, it is hard to quantify just how these contributions compare to direct editing of

[4] Data from early 2008 is not included due to some minor problems with the latest entries in the February data dump.

the encyclopedia itself. Furthermore, most of the contributions to Wikipedia occur in the article namespace, and many of the participants in other projects are also article contributors. Although additional research on activity in these namespaces is certainly warranted, measures of contribution are limited to the article namespace in order to simplify the analysis.

The final sample of users chosen for this study encompasses a set of fairly active Wikipedia editors who also have some minimal level of activity in the user talk namespace. The resulting sample consists of approximately 175,000 users. The mean activity duration for this sample, measured as the time between a user's first and last edits in the data set, is 200 days, although the median duration is less than two months.

This unusually long average activity duration highlights the major weakness with the sampling methodology used. Since users active in the user talk namespace tend to be among the most active Wikipedia contributors, the data has a tendency towards including relatively committed members of Wikipedia. While this effect is mitigated somewhat by including users who receive messages in the user talk namespace without contributing messages of their own, the potential for bias still exists. While a random sample of users is a viable alternative, this sampling approach should result in more meaningful results, as a random sample will include a large number of users who edit Wikipedia only once or twice and never return.[5] With a system as large as Wikipedia, and an attrition rate as high as the one observed in this data, finding users who remain active long enough to generate some variance in their behavior is critical to any analysis attempting to assess factors contributing to changes in a user's rate of participation. Furthermore, this sampling method helps to ensure a sufficient number of users in the sample have measurable social interactions within the site. Although this sampling approach is reasonable under the circumstances, it is

[5] Even with the restrictions employed here over 30% of the users in the sample used in this analysis edited Wikipedia on fewer than three occasions during the observation window.

important to be mindful of these potential limitations when interpreting the results of the analysis.

The time window chosen for analysis might also impact sample selection. In this case any early adopters included in the sample will represent a highly committed set of contributors, as nearly three years have elapsed between the creation of Wikipedia and the start of the data set. For this reason, no analysis comparing early adopters to other users is included in this research.

Network Data

The social interaction metrics used in this analysis are derived from interactions on user talk pages. In the raw network data, each node represents a user account and each edge represents an edit by one user on another user's talk page. Self-loops and isolates are not included in the network – self-loops because they do not add much additional information, and isolates because they are by definition not part of the original sample of users. Although the sampling methodology excludes anonymous contributors, these nodes are still included in the network if they are connected to a user in the analysis sample. These nodes are treated in the same way as nodes representing registered users. Since interactions between registered users and anonymous users are likely to be sparse and rarely repeated, creating a special category for these nodes is not necessary.

All nodes and edges existing in the user talk network from mid-2004 through the end of 2007 were extracted and stored as a directed edge list with time stamps for interactions. The resulting network consists of 1,645,323 nodes and 4,886,044 distinct edges. Although this network is extremely sparse, it has a fairly high average clustering coefficient of 0.0192. For reference, a random network with the same

number of nodes and a similar number of edges per node would have a clustering coefficient of approximately 0.0000018.

Time Periods

In order to capture changes in user behavior over time, it is necessary to split the data into time periods and measure social interaction and levels of contribution within each time window. The time periods used in this analysis are all based on the date the user first registers for an account on Wikipedia. For a given user, the first time period starts on the account creation date and extends for 84 days (12 weeks). Each subsequent time period starts the day after the previous time period ended, and extends for an additional 84 days (12 weeks). The user's final time period is determined by his/her last date of activity in the article namespace.

Activity and social interaction measures are included in a given time period based on when the action took place relative to the account creation date for each user. For example, if Bob sends Alice a message 18 weeks after he creates his account, then that message counts as an outward tie in Bob's second time period. However, if Alice created her account just 6 weeks before Bob sent the message, then it counts as an inward tie in Alice's first time period.

The length of the time periods was chosen to maximize the likelihood that a given user will have some activity during a particular time period while keeping the time periods as short as possible. This is based on empirical examination of user activity patterns on Wikipedia. Time periods varying in length from 2 to 20 weeks were created for each user, with weeks chosen as the measurement unit in order to smooth out any weekday periodicity in editing behavior. The average percentage of eligible users – defined as users active before and after the time period in question – who were active in a time period of a given length was measured and compared with

other possible time periods. A fairly sharp increase in the percentage of eligible users active during the time period was observed between 6 and 10 weeks with diminishing returns occurring at around the12 week mark. Therefore, 12 weeks was chosen as the length of the time period for this analysis.

Variables

Several network and activity metrics are derived from the data and measured for each user within each time period between the user's first and last date of activity. Two primary activity metrics appear in the final analysis. The first measures the number of days a given user edited Wikipedia during a given time period, while the second measures the difference between the number of days a given user edited Wikipedia during a given time period and the number of edit days for the subsequent time period. The first metric provides a good signal for regularity of participation, and is a reasonable measure of commitment to Wikipedia. The second indicates change in participation rate from one time period to the next. Positive numbers indicate an increase in participation, while negative numbers denote a decrease.

Edit days is used instead of raw edit counts for several reasons. First, edit days is a bounded measure. This makes it a bit easier to model, as the number of edits has an extremely long tail of heavily active contributors who skew all of the results. Second, this is a better measure of commitment. A user who edits Wikipedia every day is probably more committed – and therefore more likely to maintain activity – than a user who makes a lot of edits on one day and does not return to the site within a given time period. Finally, this measure is still fairly robust as a metric for raw contribution. Although there are users who make rare but substantial contributions to Wikipedia, over time the most consistent contributors end up producing more content and providing a greater benefit to the encyclopedia.

TABLE 3.1. Means and Standard Deviations for Wikipedia Activity Measures

Variable	mean	sd
Change in Edit Days in next time period	-1.195	11.491
Number of edit days	9.298	15.938
Inward ties from more active users	0.781	2.416
Inward ties from less active users	0.677	4.388
Outward ties to more active users	0.52	14.284
Outward ties to less active users	4.395	54.671
Mutual ties with more active users	0.457	1.959
Mutual ties with less active users	0.493	3.983
Avg inward tie strength from more active users	0.502	0.926
Avg inward tie strength from less active users	0.247	0.856
Avg outward tie strength to more active users	0.256	0.875
Avg outward tie strength to less active users	0.211	0.876
Avg mutual tie strength with more active users	0.183	0.387
Avg mutual tie strength with less active users	0.081	0.273
Closed triads with inward ties from more active users	0.102	2.283
Closed triads with inward ties from less active users	0.117	3.791
Closed triads with outward ties to more active users	0.17	31.385
Closed triads with outward ties to less active users	0.144	4.506
Closed triads with mutual ties to more active users	0.168	3.189
Closed triads with mutual ties to less active users	0.199	4.98
Avg strength within triads, inward ties from more active users	0.14	1.074
Avg strength within triads, inward ties from less active users	0.091	1.081
Avg strength within triads, outward ties to more active users	0.0726	1.139
Avg strength within triads, outward ties to less active users	0.072	0.896
Avg strength within triads, mutual ties with more active users	0.497	4.233
Avg strength within triads, mutual ties with less active users	0.318	3.398

Two types of egocentric network measures are also included in the analysis. The first type is a set of basic degree measurements: in degree, or messages sent from others to a given user; out degree, or messages sent by a user to others; and mutual ties, which represent reciprocated communication between two users. In this case,

mutual ties are excluded from in degree and out degree, so a user's total degree is equal to the sum of mutual ties, in degree, and out degree. In all cases, these degree measures are grouped by relative activity. All users have network measures for ties to alters who are more active and ties to alters who are less active than the ego user. The second type of egocentric measure is a simple measurement of average tie strength. Average tie strength is the mean weight per tie, where the weight is the number of messages sent between nodes. This average tie strength is computed for each degree measure. For example, a user receiving 9 total messages from 3 senders with a higher activity level would have an average tie strength of 3 for inward ties from more active alters.

In addition to the egocentric measures, two analogous local structural measures are included in the analysis. The first is the number of closed triads in an individual's ego network. That is, the number of pairs of a user's connections who are also connected to each other. As with the degree measures, these metrics are divided between edge type – incoming tie, outgoing tie, or mutual tie – and relative activity. For example, if a user I has mutual connections to two users, J and K, and J and K are both more active than I, then user I's triadic closure metric will increase if J and K are connected. In order to simplify processing, the IJ connection and the IK connection must both be mutual, but the JK connection is allowed to take any form. Note that triads where J sent I a message and I sent K a message are not included, nor are triads where J is more active than I but K is less active. A separate model with a more relaxed definition of triadic closure produces qualitatively similar results to the models presented here.

The second local measure is analogous to the egocentric average strength measure. This measure provides the mean number of messages per triad for each triad type. For a user I connected to nodes J and K, the triad strength is represented as the

sum of the number of messages sent between I and J and the number of messages sent between I and K. These totals are averaged across all triads to produce the final triad strength measures for inward ties, outward ties, and mutual ties in both the more and less active cases.

Table 3.1 shows descriptive statistics for all of the activity measures. Of particular note is the tendency for outgoing communication to be higher frequency actions with higher variance. Note also the change in edit days from one time period to the next is slightly negative on average. This is likely due to the relatively high probability users have of exiting the system and never returning. Finally, although the standard deviations do not always indicate such, most of the activity and network metrics are heavily skewed and follow power law or log-normal distributions.

Model Description

The time series metrics described above are included in a fixed-effects logistic regression model to determine the relationship between social interaction and subsequent change in behavior. The dependent variable in the analysis is change in edit days in the subsequent time period. The independent variables in the complete model include all egocentric and local network measures. The natural logarithm of all independent variables is used in the model in order to account for the skewness of these variables and produce normalized predictors. A control for the number of days a user edited Wikipedia during the time period under analysis is also included in each model, along with controls for both relative and absolute time periods. The relative measure is simply the user's time period. This controls for the user's familiarity with Wikipedia and longevity in the system. The absolute measure is a series of binary variables for different years, one each for 2005, 2006, and 2007, with 2004 as the control state. These variables help account for temporal effects, such as the size of

Wikipedia or changes to the interface. While year-long time periods are fairly long, even relative to the 12-week user time periods in use in this model, these controls represent a reasonable compromise between controlling for temporal shifts and keeping the number of time variables in the model reasonable.

The fixed-effects model is used because it explores variance in behavior within users across different time periods. This model is appropriate because it naturally controls for inherent user tendencies, such as native affinity or enthusiasm for Wikipedia, computer literacy, or other hidden attributes. This also helps mitigate the lack of demographic data. Ideally, a mixed effects model would be used in these circumstances in order to measure both the variance within users via a set of fixed effects and the variance between users via a set of random effects. Although such random effects models are more efficient than fixed effects models, they are also less conservative, and a Hausmann test on the data suggested that the coefficients in the two models were significantly different and therefore the random effects model could not be trusted to return valid results.

In order to assess the relative effects of tie strength, triadic closure, and triadic strength five models are estimated and assessed. The first includes controls plus the simple degree measures – in degree, out degree, and mutual connections to more and less active users. The distinction between in, out, and mutual ties is important here because in degree should indicate a more direct line of influence than out degree, but mutual ties may represent a slightly stronger connection. The second model includes the tie strength measures, which help control for variation within different types of connections. If mutual ties are important only because they are stronger and represent more frequent interaction, this model will highlight that difference. Taken together, these two models determine the extent to which dyadic relationships impact continued participation in Wikipedia.

TABLE 3.2. Results of Fixed-Effects Regression Analysis of Wikipedia Data

	Model 1	Model 2	Model 3	Model 4	Model 5
Time Period	-0.278***	-0.275***	-0.271***	-0.269***	-0.266***
	(0.0174)	(0.0174)	(0.0174)	(0.0174)	(0.0174)
Year 2005	1.884***	1.886***	1.855***	1.852***	1.834***
	(0.103)	(0.103)	(0.103)	(0.103)	(0.103)
Year 2006	3.141***	3.142***	3.091***	3.085***	3.054***
	(0.141)	(0.140)	(0.141)	(0.140)	(0.140)
Year 2007	-0.923***	-0.919***	-0.935***	-0.920***	-0.937***
	(0.171)	(0.171)	(0.171)	(0.171)	(0.171)
Edit Days	-0.617***	-0.619***	-0.622***	-0.622***	-0.621***
	(0.00435)	(0.00439)	(0.00438)	(0.00440)	(0.00441)
Inward ties from more active users	-1.146***	-1.737***	-1.077***	-1.622***	-1.632***
	(0.0583)	(0.118)	(0.0591)	(0.125)	(0.125)
Inward ties from less active users	-2.084***	-2.023***	-1.534***	-0.932***	-0.859***
	(0.101)	(0.168)	(0.104)	(0.180)	(0.180)
Outward ties to more active users	0.442***	0.547***	0.373***	0.515***	0.528***
	(0.0906)	(0.152)	(0.0954)	(0.169)	(0.169)
Outward ties to less active users	-0.235***	-0.425***	-0.215***	-0.436***	-0.418***
	(0.0779)	(0.101)	(0.0778)	(0.102)	(0.102)

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

TABLE 3.2 (Continued)

	Model 1	Model 2	Model 3	Model 4	Model 5
Mutual ties with more active users	0.00232	0.487***	-0.0178	0.620***	0.607***
	(0.0876)	(0.172)	(0.0930)	(0.195)	(0.195)
Mutual ties with less active users	-0.105	0.0403	0.0440	0.282	0.473*
	(0.138)	(0.214)	(0.150)	(0.246)	(0.249)
Avg inward tie strength from more active users		0.833***		0.734***	0.755***
		(0.123)		(0.127)	(0.128)
Avg inward tie strength from less active users		-0.00339		-0.838***	-0.791***
		(0.173)		(0.180)	(0.180)
Avg outward tie strength to more active users		-0.166		-0.166	-0.189
		(0.148)		(0.156)	(0.156)
Avg outward tie strength to less active users		0.578***		0.555***	0.554***
		(0.144)		(0.144)	(0.144)
Avg mutual tie strength with more active users		-0.358***		-0.416***	-0.402***
		(0.0967)		(0.104)	(0.103)
Avg mutual tie strength with less active users		-0.137		-0.297**	-0.305**
		(0.139)		(0.150)	(0.151)
Closed triads with inward ties from more active users			-0.245	-0.0487	0.346
			(0.174)	(0.180)	(0.308)
Closed triads with inward ties from less active users			-3.409***	-3.654***	-1.765***
			(0.258)	(0.268)	(0.386)

TABLE 3.2 (Continued)

	Model 1	Model 2	Model 3	Model 4	Model 5
Closed triads with outward ties to more active users			-0.137	-0.225	-0.457
			(0.228)	(0.240)	(0.340)
Closed triads with outward ties to less active users			0.466*	0.451*	0.536
			(0.246)	(0.248)	(0.342)
Closed triads with mutual ties to more active users			-0.111	-0.371*	-0.171
			(0.182)	(0.195)	(0.291)
Closed triads with mutual ties to less active users			0.803***	0.667***	1.084***
			(0.240)	(0.251)	(0.340)
Avg strength within triads, inward ties from more active users					-0.369*
					(0.206)
Avg strength within triads, inward ties from less active users					-2.241***
					(0.297)
Avg strength within triads, outward ties to more active users					0.303
					(0.267)
Avg strength within triads, outward ties to less active users					-0.455
					(0.299)
Avg strength within triads, mutual ties with more active users					-0.117
					(0.137)
Avg strength within triads, mutual ties with less active users					-0.526***
					(0.186)

TABLE 3.2 (Continued)

	Model 1	Model 2	Model 3	Model 4	Model 5
Constant	5.932***	5.894***	5.925***	5.913***	5.911***
	(0.105)	(0.105)	(0.105)	(0.105)	(0.105)
Observations	368188	368188	368188	368188	368188
Number of userid	80931	80931	80931	80931	80931
R-squared	0.365	0.365	0.366	0.367	0.368

The third model includes all controls and degree measures along with triadic closure counts. This model assesses the importance of triadic closure as a predictor of changes in contribution. Based on the findings in Backstrom et al. (2006), triadic closure is expected to predict a shift towards the friend's levels of activity. That is, users with closed triads involving less active users should decrease activity, while users with closed triads involving more active users should increase activity. The last two models are designed to assess whether or not any triadic closure effects are a function of tie strength. They add controls for dyadic tie strength and average tie strength within triads. Table 3.2 shows the results of the regression models.

Discussion of Results

Based on previous work on social relationships and continued participation, an increase in dyadic ties to more active users should be correlated with a subsequent increase in individual activity while an increase in ties to less active users should be correlated with a decrease in subsequent contribution. Furthermore, stronger ties should have a stronger effect, and therefore the effects of tie strength measures should follow the same pattern. These predictions are tested in models 1 and 2, with mixed results.

Although the effects of increasing outward ties are in line with theoretical expectations, the effects for incoming ties are consistently negative. Furthermore, the effects of mutual ties are not statistically significant. Tie strength measures produce similarly mixed signals. While increased strength of incoming ties from more active users has the expected direct effect, including this measure in the model causes the effect of additional ties to become more strongly negative, suggesting that inward ties are generally associated with a subsequent decrease in participation. Conversely, outward tie strength with less active users is positively correlated with subsequent

participation, but including this measure strengthens the negative relationship between outbound ties to less active users and subsequent participation. Similarly, increasing the strength of mutual ties to more active neighbors predicts a subsequent decline in participation, but increasing the number of such ties becomes positive and significant. While most of these results support earlier findings that dyadic social ties are useful predictors of participation, the measure with the strongest effect actually runs counter to expectations. Regardless of controls for tie strength, incoming ties remain strongly and negatively correlated with subsequent participation regardless of the relative activity levels of the focal actor and his/her neighbors.

One possible explanation for the consistently negative effect of incoming ties is a sanctioning effect, where more experienced users send messages over the user talk network chastising newer users for poor behavior. It is possible that a large proportion of users who receive sanctioning messages respond by dropping out of Wikipedia rather than responding. Another explanation is related to the probability that unresponsive users have already left the system. Such users may receive messages within the system but never actually read them, leading to a spurious effect where inward ties predict subsequent decline in participation even though the decline in participation was unaffected by these messages.

Model 3 tests the findings presented by Backstrom et al. (2006) around group joining behavior, where increased triadic closure is expected to cause the individual to behave more like the peer group. Based on their findings, triadic closure should correlate with an individual subsequently changing behavior to correspond more closely to the actions of his or her peers. In line with these expectations, increasing triadic closure amongst in-degree neighbors with less activity predicts a subsequent decrease in user activity. However, due to the nature of unreciprocated incoming ties it is possible that this finding does not indicate any particular importance for triadic

closure. Indeed, these data are perfectly aligned with both of the explanations offered for the persistent negative relationship between unreciprocated incoming ties and subsequent contribution.

The other statistically significant effect associated with triadic closure is a positive relationship between mutual triads with less active neighbors and a subsequent increase in participation. This result is slightly puzzling, particularly in the absence of other significant effects for similar metrics. One possible explanation for this effect of mutual triads is some form of social affirmation. Even though the others involved in the exchange are less active, the fact that they are there, connected, and contributing at some level confirms the value of the contribution in the mind of the user, leading to a small increase in subsequent contributions to Wikipedia.

The addition of tie strength variables in models 4 and 5 reveal two interesting effects. First, the effect of triadic closure for inward ties is modified by the strength of those triads under certain circumstances. When a measure of triadic tie strength is included, the primary effect of triadic closure amongst incoming neighbors is also reduced, suggesting that this particular measure of local clustering is encoding some of the same effects as dyadic tie strength. Furthermore, when a user is receiving lots of contact from two connected users who are less active, s/he will be less likely to continue editing. This effect fits with earlier explanations for the general effect of incoming ties. Users who have dropped out will never respond, no matter how many messages they receive, and users who reduce their contributions due to sanctioning may be more likely to do so in the face of repeated sanctioning.

Second, stronger ties within mutual triads where both neighbors are less active are correlated with a decrease in subsequent contribution. However, the correlation between the number of such triads and a subsequent increase in contribution is strengthened when measures of triadic tie strength are included in the model,

suggesting that in this case triadic closure is capturing a separate mechanism from the effect of dyadic tie strength. For example, social influence may be operating through tie strength as users become more similar to their closest friends. Meanwhile an increasingly cohesive social group may provide stronger social affirmation, causing an individual to increase contribution even though her friends are less active participants. Alternatively, tie strength may indicate an increase in social activity in Wikipedia at the expense of contributions to the encyclopedia. If users have a limited amount of time to spend editing Wikipedia, then there will be a natural tension between their contributions to the encyclopedia and their time spent socializing. A social area of the site that fosters a community may well be important for encouraging continued participation, but if it becomes enough of a distraction the social aspect could ultimately hurt the production of the encyclopedia.

The impact of the control variables remained fairly consistent across all 5 models. Users editing Wikipedia in 2005 and 2006 were more likely to increase their activity over time than users editing wikipedia in 2004, while users editing Wikipedia in 2007 were more likely to decrease their activity over time than users editing in 2004. This may be due to interface or changes to the underlying user base.

Generally speaking, longer-term users were more likely to decrease activity. This may be due to generally high levels of activity within this population, or a generally higher probability of moving on to other activities within Wikipedia. This effect is fairly small, and may in fact be smaller than expected given that users are probably more likely to leave Wikipedia altogether after some length of time.

Activity in the previous time period is negatively correlated with a user's subsequent change in contribution. This makes intuitive sense – at a certain point, it is no longer possible to increase contributions. This is particularly true when the measure is number of edit days, which has a maximum of 84 for any given time period.

Furthermore, users with a high activity rate will have a larger negative delta when they leave Wikipedia altogether. This likely skews the mean values for the dependent variable somewhat lower.

Summary of Findings

Based on the models presented here, direct social connections can serve as predictors of participation, but in a system like Wikipedia where contributions are low-cost the need for a social incentive may be low enough that the effects of social connections are smaller than those observed in other settings. Although there are mixed results for the importance of dyadic relationships, local clustering amongst particular interest groups has an effect, and social groups may influence individuals to adjust their levels of contribution. In other words, having a cluster of connected friends in the system may lead to social affirmation of the inherent value of making contributions, which then leads to an increase in contribution. Furthermore, triadic closure and tie strength appear to have different effects, although under certain circumstances some of the impact of tie strength is encoded by measures of triadic closure.

Although the Wikipedia data generally supports the notion that social relationships will predict subsequent behavior, the specific results presented here occasionally vary from the theoretically predicted relationships. While some of the deviation from the expected results may be caused by the nature of unreciprocated ties, or the operation of some form of social affirmation through local clustering and reinforcement, it may well be the case that social connections are not as important in low-risk, low-cost activities as they are in higher risk activism. On the other hand, the mixed results may be due to the nature of the user talk space itself. Some of the ties are not purely social, and sometimes this space is used for Wikipedia-related

discussion that does not fit anywhere else on the site. Furthermore, although the network data for the user talk namespace is excellent, it is still missing other modes of communication Wikipedia users might adopt, such as email or face-to-face interaction. In a service with a relatively poor social interaction component, these external communications media might play a larger role in the community formation and social reinforcement that takes place amongst contributors than they would in an inherently social platform.

The next chapter presents an analysis of Wallop, a blogging and social networking service. This analysis will focus on a smaller network with a more clearly social orientation and an emphasis on communication and interaction over production of a public good. The results presented in the following chapter may shed some light on which of the effects observed in the Wikipedia analysis are due to something specific to the service and which are more fundamental to the way individuals decide t contribute to a public good.

CHAPTER 4:
ANALYSIS OF WALLOP PARTICIPATION

Wallop Analysis

The analysis presented in Chapter 3 suggests that connections to a social group are important predictors of contribution to Wikipedia. However, these results are somewhat mixed, and they may be due to some particular element of Wikipedia's design. Therefore, the second stage of this research replicates the Wikipedia analysis on data collected from Wallop, a weblogging and social networking system.

The Wallop data complements the Wikipedia data nicely. Like the Wikipedia data, it includes a time-stamped log of user activity on the site, complete with data on social interactions. Unlike Wikipedia, the social interaction data are not cleanly divided from data on contributions to the public good. Wikipedia is designed to accomplish a specific task – writing an encyclopedia – and social interaction within Wikipedia takes place in a separate section of the site. Wallop is explicitly designed to encourage an interconnected group of users who interact with one another around the content they publish. There is no distinct public good in Wallop, because the public good is the community itself. Any content production – whether original and unsolicited or created as part of an interaction with another user – constitutes a contribution to the Wallop community. Therefore, the Wallop data has a stronger social network component than Wikipedia, and the social interactions in the Wallop data are inextricably linked with the data on content contribution.

The analysis of the Wallop data mirrors the Wikipedia analysis in order to assess similar questions in a different setting. This analysis explores the impact of social relationships on participation in a situation where social interaction is of central importance, and the concept of a public good is not clearly defined. This provides a

natural comparison of the importance of social relationships as a predictor of contributions to a task-oriented system such as Wikipedia with the importance of such relationships as a predictor of participation in a socially oriented system like Wallop. It also makes it possible to assess the robustness of any effects seen in both settings.

The remainder of this chapter focuses on the analysis of the Wallop data. It begins by describing the Wallop weblogging system and moves to a detailed discussion of the data collection, processing, and network data extraction used in this study. This is followed by a description of the variables and analytical methods used in this analysis. The chapter concludes with a discussion of the results of this study, and a comparison of the findings with the results from the Wikipedia analysis.

About Wallop

While it was active, Wallop was a personal publishing and social networking system. Built by Microsoft Research and launched as a research prototype in late 2004, it was designed to facilitate social interaction through a variety of interface elements. Following the prototype launch the system expanded for a period of about 8 months before the development team moved on to other projects. Eventually the prototype version was disabled, although the technology behind the service lived on for a time in the form of a third-party social networking system.

Although one of the research goals of the system was to facilitate interaction, it was still primarily designed as a personal publishing, or blogging, service. The default interface was a personal blog, where users could write text-based posts or upload multimedia. Other users could then navigate to this page and leave comments directly on any piece of content added to the blog space. Creating text-based posts, uploading photos, and commenting on content were the most common activities in Wallop.

Wallop also included a visible social networking interface. This enabled users to connect with one another and find each other's content quickly and easily. The primary component of this interface was the personal social network map. This appeared as a section to the left of the blog. The area immediately visible on the left-hand side of the screen was limited and could only contain 15 other user names, but a link allowed it to expand into the blogging window to show an additional 35 names. A user could add someone to his/her network by dragging that person's name into the social network map. A user could then navigate to that person's blog by clicking on the appropriate name in the social map. From there, the user could see that person's social map, and click through to a third user's page, and so on. This allowed users to browse through connections by following links from one user page to another.

For new users, or those who did not have many people in their social networks, the system would recommend a person of interest by adding that person's name to the user's social network map. This recommendation was based on an algorithm related to the user's activity and interactions with others on the site. If the user already had several connections the system would not add any additional recommendations.

Although it was not visualized, Wallop also allowed users to navigate the invitation hierarchy of the site by viewing the inviter for every other user on the site. This was possible in part because Wallop was an invitation-only service. Users had to be invited by other users before they could join, and it was not possible to walk up to the site and register for an account. Therefore, the initial set of users were all invited by developers and testers involved in the project. They invited their friends, who invited their friends, and so on over the course of the next 12 months. By the end of its first year, over 135,000 users had been invited to join Wallop, and over 80,000 had logged in at least once.

Despite expanding strictly through invitations from a small set of American software developers, Wallop was populated by users from all over the world. Users logged in to Wallop from numerous countries, and several languages were represented. Indeed, the dominant language in the service was Chinese, as over 35,000 of the approximately 60,000 content contributors used Chinese on a regular basis (Gu et al., 2006).

Content Contribution and Social Interaction

Users contributed original content to Wallop by making posts through the blogging interface. These posts covered a broad range of topics, from technology to fashion to day-to-day activities, and could include any combination of photos, music, and text. There was no limit to the number of blog posts a user could make, but music uploads were limited to 50 songs, and photo contributions were capped at 5000. Although several users reached the music upload limit, it was uncommon for any user to reach the maximum number of photos.

In addition to posting content to their own blogs, users could interact with other content on the site. They could comment on blog posts, photos, music, other comments, and any other visible content on the site. Users could, and often did, comment on their own content, but many of the comments were attached to content produced by others. The ability to comment on the comments left by other users frequently led to threaded conversations, where users would reply back and forth to one another rather than directing all comments at the original content creator. By attaching comments to content produced by others Wallop users could engage in direct interaction with one another. In addition to exchanging public comments, users could also send private comments. These would trigger an email to the content creator, which s/he could then respond to either publicly or privately.

Indirect interaction was also possible through the Wallop interface. The most common form of indirect interaction involved a user adding a target to his/her network map. Inclusion in a user's network map generally indicated some relationship between the user and the target, although the nature of that relationship was not always clear. In some cases, it was an indication that the target user was close to the user, while in other cases it was a simple bookmarking activity so a user could quickly navigate to a target known to post interesting content or particularly nice photos. These interactions were not nearly as common as comment interactions, partly because the network map was not especially dynamic. They were also less public – although not technically hidden, there were fewer navigation paths to a given user's network map, and therefore the visibility of connections between users was lower than the visibility of connections between content and comments.

One final mode of indirect interaction was photo tagging. Users could mark regions of any visible photo on the site, and associate comments or users with those regions. In this way, it was possible for users to tag others in a photo. These photo tags indicate a high probability of a face-to-face relationship between a user and a target, or between two users tagged in the same photo. However, photo-tagging events were rare in Wallop, and tended to occur among the most active users.

Data and Methods

Since Wallop was a proprietary piece of technology the data are not publicly available. Fortunately, Microsoft Research provided access to a snapshot of the primary back-end database for the Wallop system. This database contains information on all interface elements in Wallop, including user account data, time-stamped content creation logs including photo and music uploads, blog posts, and comments. The database includes all of the metadata for user activity on the site along with full text of

comments and blog posts, but multimedia content was originally stored in a separate database and is therefore not available for analysis.

In addition to the raw activity logs, the Wallop database also includes an association data table. This table contains a record of all the links between objects on the site. These objects can be people, photos, comments, or blogs, and they are associated by connections in the interface. For example, a comment on a photo would be listed in the association table as a link between the photo and the comment.

Since the data are stored in a SQL database, pre-processing of XML and text data was not necessary for this stage of the analysis. However, the database was originally designed to drive the Wallop site, not to facilitate research. Therefore, several preliminary queries were used to re-format the data into action and network data tables matching the SQL tables used in the Wikipedia analysis. Once these preliminary queries were executed the same processing scripts used in the Wikipedia analysis could be applied to the Wallop data. This resulted in a data set ready for processing by the same R and Stata scripts used to conduct the analysis on the Wikipedia data.

Deriving Social Network Data from Wallop

As with Wikipedia, there are several ways to measure social ties in the Wallop system. Using the invitation tree is arguably the most direct approach. This method traces invitations from sender to recipient. The nodes in the invite graph represent Wallop users, and the edges represent invitations to join Wallop. This approach focuses on direct connections between users, as sending an invitation requires the sender to know the recipient's email address. However, interaction in the invitation network can only move in one direction, and apart from rare cases where a user closes an account and receives a second invitation the interactions can only happen once,

resulting in a mostly static network. Furthermore, users could only send a limited number of invitations, so the range of variation in the number of ties for a given user is not particularly large. Although valuable for some applications, measuring the invitation graph presents a heavily restricted view of Wallop's underlying social network.

A better alternative might be to use network map data. Users can add network connections by dragging others into the network map space. In this graph, nodes are users and edges represent whether or not a given target node appears in another user's network map. This data represents the social network presented to the user with an explicit visualization in the interface, and users are more likely to directly manage this network than any other on the site. Despite the obvious potential for users to adjust this network, the network map data are largely static. Users did not change their network maps all that often, and rarely removed other users from their networks. Furthermore, network measures based on this interface do not capture any direct interaction between users. Indeed, it is not entirely clear what these metrics capture – it might be capturing bookmarking behavior, affinity, closeness to another user, or any combination of the above.

An interaction-based network measure can be derived from the comment network, where connections are formed by interaction through comments. In this network, the nodes are users and the edges are comments sent from one user to another. This metric captures a high-frequency form of direct interaction, which is generally visible and used by a majority of the active users on the site. Since comments can be directed at other comments, edges usually represent communication between individual users. For example, if Chris wants to say something about a comment Alice left on Bob's photo, he can do so by commenting directly on Alice's content. This results in a comment thread where each comment is linked to a comment

made by a distinct user. While threading is not perfect, as users may respond to the wrong object or respond to multiple objects at once, it does provide a great deal of structure to the interactions on the site. Furthermore, users do typically attach comments to the target of their response. It is therefore reasonable to assume that a given comment will be directed at the owner of the particular piece of content to which it is attached. The major drawback of this measure is that comments are fairly cheap. That is, it is easy to leave a comment for another user, so the fact that two users are connected by a single comment may be less meaningful than the fact that two users are linked by an invitation or a mutual network map connection. However, this limitation is mitigated by the fact that the lighter weight commenting actions are more similar to editing a user's talk page in Wikipedia than either sending an invitation or adding the user to the network map.

One final set of network measures can be derived from photo tagging and co-occurrence in photos. In this network, nodes are users and edges could be either the act of tagging another user in a photo, the co-occurrence of users in a photo, or both. In this case, an edge would exist between Alice and Bob if Alice tags Bob in a photo, or if Alice and Bob are tagged in the same photo. This measure presents a strong indication that two users know each other face-to-face, and in the case of co-occurrence suggests a certain level of shared offline activities. Unfortunately, photo tagging is a low-incidence event, and misses the bulk of the online social interaction that takes place on the site.

This analysis defines network edges as direct communication through comments. This network representation is most similar to the one used in the Wikipedia analysis, and it is based on arguably the best representation of social interaction within Wallop. Although photo tags or invitations might add additional weight to a comment, the relative lack of such edges makes using them impractical,

and the static nature of invitation and network map edges makes them less useful as predictors of changes in participation rates over time. By contrast, comments are common and dynamic, and most active users engage in some degree of commenting behavior during their time on the site.

Sampling

Although Wallop is a much smaller system than Wikipedia, the raw data logs still represent an enormous amount of information. In addition to the accounts of the 130,000+ users invited to join the system, the database also includes data on every action from the roughly 80,000 users who eventually responded to the invitation and registered their accounts. This includes time stamps, user ids, object ids, and additional metadata for millions of comments, photos, and other objects created on the site. It also includes time stamps and object metadata for millions of associations, or links between two connected objects, on the site.

In order to avoid polluting the sample with data from users who did not use the service, this analysis only includes registered user accounts with at least some content creation activity. As with the Wikipedia analysis, users who did not either send or receive a comment are also excluded. This restriction was imposed both because the primary independent variables in this analysis are based on social interactions, and because this restriction mirrors the sampling restrictions placed on the Wikipedia data. Like participants in the Wikipedia user talk network, Wallop users involved in the comment network tend to be more active contributors to the system than those who never send or receive a single comment. Unlike Wikipedia, where only a small fraction of users have any activity in the user talk network, nearly 40% of Wallop users had some activity in the comment network during the observation period. Considering that nearly half of Wallop users logged in 0 or 1 times during the

observation period, and that the majority of these users are not included in the comment network, this data sample represents a substantial majority of active Wallop users.

The database snapshot used in this analysis includes all data from the early stages of development through the end of August 2005. Since the early stages of development do not represent real user behavior, it is excluded from this analysis. Only data from Wallop's public release in late September 2004 through the end of the available data are included in this analysis. Furthermore, developers and administrators are excluded from the analysis, although their interactions with other users are included in the relevant network metrics.

The final sample of users consists of approximately 34,000 Wallop users with at least one connection in the comment interaction network. The mean activity duration, measured as the number of days between their first and last content contribution events, is 88 days, with a median duration of 65 days. These are fairly long activity durations given that the total observation window is just under 1 year and over half of these users joined Wallop several months after the public launch.

Like the sample selected for the Wikipedia analysis, this data sample is biased towards more active and dedicated Wallop users, but the risk of biased results must be taken in order to generate a data sample with usable network metrics. The possible effects of sampling bias are mitigated somewhat by the analytical approach, which focuses on changes in user behavior over time. Furthermore, many of the users excluded from the sample have no meaningful activity within the Wallop system so their exclusion should not have a substantial impact on the results of the models. In the Wikipedia analysis, including users in the sample who received one or more user talk messages but did not send any reduces some of the inherent bias in the sampling approach, as 48% of the users included in the Wikipedia data fit into this category. In

Wallop, this approach does not have as large an effect, as only 28% of users in the comment network have never sent a message. However, since commenting is one of the most common forms of user activity in Wallop, this sample captures a much larger percentage of the user population than the sample used in the Wikipedia analysis.

Network Data

The social network data used in this analysis is derived from interactions through comments sent either directly as private messages or posted publicly on another user's content. In the raw network data, each node represents a user account and each edge represents a comment. Self-loops are removed from the network because they do not add additional information and because the interface encourages the creation of self-loops. Indeed, the only way a user could add a caption to a photo on the Wallop site was to attach it as a comment, so many of the comments in the database were self-directed. Isolates are not included in the network data since they are not part of the sample of users under analysis.

Although they are not included in the analysis, developers and testers are included in the network data. They are not included in the analysis because they generally remained heavily active, and in some respects were extreme outliers in terms of their behavior on the site. However, during the observation period the extreme behaviors of developers and testers were primarily limited to new feature launches and content types outside of the normal scope of user activity, and in terms of their actions within the comment interaction network they were much like other users. Furthermore, their social connections to others could still be important predictors for subsequent user activity. Therefore the network data includes their activity, and the network metrics of any user connected to a developer or tester includes those connections. A small number of users who never logged in to Wallop are also included in the network

data, as they received private messages from existing users. They are not included in the subsequent analysis since all of their activity metrics are null.

The network data was constructed from the affiliations table in the Wallop database, which includes a list of connections between objects on the site. These affiliations were translated into a directed user-to-user edge list with time stamps for interactions. All edges from Wallop's public launch in late September 2004 through the end of the data sample in August 2005 are included in the network, which consists of 37,307 nodes and 156,438 edges. This network is sparse, but its average clustering coefficient of 0.12169 is extremely high when compared to the expected clustering coefficient of approximately 0.0001124 for a random graph of the same size.

Time Periods

The time periods used in this analysis are defined in a manner similar to that used in the Wikipedia analysis, but due to differences between the systems some of the specifics have been modified for this analysis. Unlike Wikipedia, where account creation requires an intentional action on the part of the user, Wallop accounts are created whenever a user is invited to join the service. Therefore, the first time period begins on the day the user first logs in to Wallop, rather than the date the account is created. Each time period is 56 days (8 weeks) long, with subsequent time periods beginning the day after the previous time period ends. The final time period is determined by the later of the user's last content upload or the user's last log in date.

Time periods are measured in weeks in order to smooth out any weekday periodicity in Wallop usage, and the length of the time periods is optimized to maximize the percentage of users engaging in activity in a given time window while minimizing the length of the time periods. This time window was chosen empirically, using an approach identical to the one employed in the Wikipedia analysis. In this

case, the point of diminishing returns in the distribution of activity over time is approximately 8 weeks.

As with Wikipedia, user activity and network metrics are all measured within a given time period relative to the user. This means if two users, Alice and Bob, interact in the Wallop comment network a given interaction might count in a different time period for each user. For example, if Alice registers and sends Bob a message during her ninth week of activity, this message will count as an outward tie during Alice's second time period. However, if Bob only joined Wallop a few days prior to the message, it would count as an inward tie for Bob during his first time period. This approach ensures that each predictor is measured within the appropriate time period for a given user.

Variables

The variables used in this analysis match the variables used in the Wikipedia analysis as closely as possible. The primary activity metrics used in this analysis are the number of activity days within a given time period and the change in the number of activity days from one time period to the next. An activity day is defined as a day where the user contributed content to the system in the form of a blog post, multimedia upload, or comment, including comments directed at other users.

Unlike in the Wikipedia context, where social interaction is cleanly divided from contributions to the encyclopedia, directed comments in Wallop are all part of the same system. Furthermore, content produced through interaction with other users is part of the public good provided in Wallop – indeed, the collective benefit of Wallop as a system is the community and the interactions around content produced by members of that community. Since these activity measures are designed to track user contributions to the public good, and since comments add value to the system as a

whole, it makes sense to include social interactions in this measure. Including comments as part of the activity metrics tends to increase the correlation between user activity in a given time period and the egocentric network metrics associated with that user. However, this increased correlation does not represent major difficulty for the models since directed comments represent a fraction of the total activity for most users. There is also some separation between the network metrics in a given time period and the change in user activity into the subsequent time period. Furthermore, this problem is not present for triadic closure metrics since the presence of a closed triad is dependent on the actions of someone other than the focal user.

The network metrics computed for this analysis are analogous to those computed for the Wikipedia analysis. For each user, in degree, out degree, and mutual ties are recorded for a given time period. Average tie strength, defined as the total number of comments divided by the number of edges, is also computed for in, out, and mutual connections. These measures are divided into two groups: connections to users with higher levels of activity and connections to users with lower levels of activity during the given time period. Triadic closure metrics are also computed for triads where both alters are more active and triads where both alters are less active. In order to differentiate between inward, outward, and mutual ties triad counts are divided into sets where both alters are tied to the user by the same type of connection. A measure of average strength, defined as the number of messages between the focal actor and a neighbor in a closed triad divided by the number of triads, is also computed for each type of triangle in a user's local network.

Table 4.1 shows descriptive statistics for these measures. The means and standard deviations are computed across all time periods. As with the Wikipedia analysis, the average change in activity days from one time period to the next is negative, likely caused by a high probability of a user leaving the system and never

returning. The network stats for Wallop are also interesting. Of particular note are the low average values for one-way edges. Indeed, there are no cases where a user has an incoming tie from a less active user, and the edge count, tie strength, and triad measures for mutual ties are all higher than the same metrics for one-way edges. This suggests a higher degree of reciprocation and social interaction in Wallop than that observed in Wikipedia.

Another important trend in these data are the tendency for users to have more mutual ties to less active users than mutual ties to more active users. This is probably caused by skew in the data. A few heavily active users might be connected to many relatively inactive alters, while less active users are only going to be connected to a few more active alters. This tends to inflate the mean for connections to less active users. More generally, most of these measures are heavily skewed, and fit either a power-law or log-normal distribution.

Model Description

A fixed-effects logistic regression model is used to determine the relationship between social interaction and changes in participation rate. The dependent variable is change in activity days in the subsequent time period, while the natural logarithms of the network metrics from the current time period are used as independent variables. A control for user activity, measured as user activity days in the current time period, is included in the model. The user time period is also used to control for the passage of time relative to the user's first experience with Wallop, while a set of dummy variables are used to measure the absolute time in which an event takes place. The observation window is divided into four time periods of three months each, with the first three months used as the reference category and dummy variables for the other time periods included as controls. As with the Wikipedia analysis, these temporal

controls help smooth out the effects of a user's familiarity with the system and any changes to the system itself.

TABLE 4.1. Means and Standard Deviations for Wallop Activity Measures

Variable	mean	sd
Change in Activity Days in next time period	-2.729	5.337
Number of activity days	3.997	6.678
Inward ties from more active users	0	0
Inward ties from less active users	0.582	1.806
Outward ties to more active users	0.103	0.561
Outward ties to less active users	0.523	1.94
Mutual ties with more active users	0.273	0.804
Mutual ties with less active users	0.725	2.979
Avg inward tie strength from more active users	0	0
Avg inward tie strength from less active users	0.375	0.665
Avg outward tie strength to more active users	0.1	0.4
Avg outward tie strength to less active users	0.284	0.565
Avg mutual tie strength with more active users	1.493	7.07
Avg mutual tie strength with less active users	1.506	4.443
Closed triads with inward ties from more active users	0	0
Closed triads with inward ties from less active users	0.097	1.141
Closed triads with outward ties to more active users	0.018	1.657
Closed triads with outward ties to less active users	0.038	0.513
Closed triads with mutual ties to more active users	0.135	1.65
Closed triads with mutual ties to less active users	0.849	11.652
Avg strength within triads, inward ties from more active users	0	0
Avg strength within triads, inward ties from less active users	0.108	0.594
Avg strength within triads, outward ties to more active users	0.017	0.225
Avg strength within triads, outward ties to less active users	0.05	0.429
Avg strength within triads, mutual ties with more active users	0.93	7.05
Avg strength within triads, mutual ties with less active users	1.918	10.495

TABLE 4.2. Results of Fixed-Effects Regression Analysis of Wallop Data

COEFFICIENT	Model 1 depvar	Model 2 depvar	Model 3 depvar	Model 4 depvar	Model 5 depvar
Time Period	-0.889***	-0.935***	-0.955***	-0.972***	-0.971***
	(0.0470)	(0.0472)	(0.0472)	(0.0473)	(0.0472)
Months 3-6	0.261***	0.211***	0.225***	0.184**	0.181**
	(0.0782)	(0.0778)	(0.0776)	(0.0774)	(0.0773)
Months 6-9	0.304**	0.224*	0.251**	0.203*	0.199*
	(0.121)	(0.121)	(0.120)	(0.120)	(0.120)
Months 9-12	-1.638***	-1.688***	-1.642***	-1.681***	-1.685***
	(0.209)	(0.208)	(0.207)	(0.206)	(0.206)
Activity Days	-0.905***	-0.895***	-0.928***	-0.922***	-0.922***
	(0.0116)	(0.0120)	(0.0120)	(0.0128)	(0.0128)
Inward ties from more active users	0	0	0	0	0
	(0)	(0)	(0)	(0)	(0)
Inward ties from less active users	-0.948***	-0.924***	-0.885***	-0.830***	-0.830***
	(0.0684)	(0.144)	(0.0770)	(0.175)	(0.173)
Outward ties to more active users	-0.461***	0.00271	-0.391***	0.132	0.122
	(0.112)	(0.280)	(0.116)	(0.312)	(0.312)

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

TABLE 4.2 (Continued)

COEFFICIENT	Model 1 depvar	Model 2 depvar	Model 3 depvar	Model 4 depvar	Model 5 depvar
Outward ties to less active users	0.371***	1.357***	0.209**	1.131***	1.137***
	(0.0809)	(0.168)	(0.0844)	(0.183)	(0.182)
Mutual ties with more active users	-0.736***	-1.003***	-0.727***	-1.316***	-1.331***
	(0.101)	(0.232)	(0.106)	(0.271)	(0.273)
Mutual ties with less active users	-0.207**	0.279*	-0.869***	-0.648***	-0.646***
	(0.0902)	(0.157)	(0.0985)	(0.187)	(0.187)
Avg inward tie strength from more active users		0		0	0
		(0)		(0)	(0)
Avg inward tie strength from less active users		-0.107		-0.0880	-0.123
		(0.165)		(0.182)	(0.183)
Avg outward tie strength to more active users		-0.495*		-0.540*	-0.533*
		(0.264)		(0.284)	(0.284)
Avg outward tie strength to less active users		-1.742***		-1.415***	-1.428***
		(0.202)		(0.211)	(0.212)
Avg mutual tie strength with more active users		0.162		0.262**	0.276**
		(0.113)		(0.123)	(0.125)
Avg mutual tie strength with less active users		-0.527***		-0.170*	-0.190*
		(0.0915)		(0.102)	(0.103)

TABLE 4.2 (Continued)

COEFFICIENT	Model 1 depvar	Model 2 depvar	Model 3 depvar	Model 4 depvar	Model 5 depvar
Closed triads with inward ties from more active users			0	0	0
			(0)	(0)	(0)
Closed triads with inward ties from less active users			-0.268	-0.308	-0.740**
			(0.189)	(0.208)	(0.378)
Closed triads with outward ties to more active users			0.329	-0.0502	-0.122
			(0.392)	(0.437)	(0.812)
Closed triads with outward ties to less active users			0.975***	0.355	0.191
			(0.351)	(0.371)	(0.738)
Closed triads with mutual ties to more active users			0.347*	0.556**	0.715*
			(0.207)	(0.226)	(0.369)
Closed triads with mutual ties to less active users			1.234***	1.018***	0.935***
			(0.147)	(0.161)	(0.243)
Avg strength within triads, inward ties from more active users					0
					(0)
Avg strength within triads, inward ties from less active users					0.461
					(0.281)
Avg strength within triads, outward ties to more active users					0.0702
					(0.622)

TABLE 4.2 (Continued)

COEFFICIENT	Model 1 depvar	Model 2 depvar	Model 3 depvar	Model 4 depvar	Model 5 depvar
Avg strength within triads, outward ties to less active users					0.185
					(0.486)
Avg strength within triads, mutual ties with more active users					-0.0785
					(0.137)
Avg strength within triads, mutual ties with less active users					0.0726
					(0.113)
Constant	2.177***	2.413***	2.391***	2.508***	2.513***
	(0.0830)	(0.0842)	(0.0841)	(0.0835)	(0.0833)
Observations	64196	64196	64196	64196	64196
Number of userid	32148	32148	32148	32148	32148
R-squared	0.673	0.677	0.678	0.679	0.679

The five models used in this analysis are identical to the models used in the Wikipedia analysis. The first models include only egocentric network measures and relevant tie strength measures in addition to the control variables. The third model assesses the importance of triadic closure. The final two models include measures of tie strength, both within triads and across all nodes in a given user's ego network. These models test the relationship between triadic closure and tie strength as predictors of changing participation. Table 4.2 shows the results of the models, staged from a simple model containing only simple degree measures up to models including all of the independent variables of interest.

Discussion of Results

The first two models provide little support for earlier predictions regarding the correlation between social relationships and changes in contribution. The effects of inward ties from less active users are correlated with a decrease in subsequent contribution, which matches expectations regarding an individual's tendency to become more similar to his or her friends over time. However, as in the Wikipedia analysis, this may be a spurious effect that results from an individual leaving the service before receiving the incoming messages in question. Furthermore, unreciprocated ties are fairly rare in Wallop, making it more likely that this result is not indicative of a homophily or influence mechanism, but rather is capturing a correlation between leaving the system and failing to respond to incoming messages.

Outward ties defy expectations, as sending messages to more active users predicts a subsequent decrease in activity while sending messages to less active users predicts an increase in activity. Including controls for tie strength causes the results for outward ties to more active users to lose significance, and in general this measure is not robust. However, the impact of increasing the number of outward ties to less active

users becomes more strongly positive when tie strength measures are included. Meanwhile, increasing the average outward tie strength to such users predicts a decrease in participation. This combination of results suggests that social interaction does have some impact. Users sending a large number of unreciprocated messages are probably sending them to people who have left the service, and due to the lack of response such users are themselves reducing their participation. Conversely, users who send messages to a large number of recipients are more likely to receive a response in the subsequent time period, which could itself cause increased participation. While this is not the same mechanism proposed in earlier theoretical and empirical literature, it does point to social interaction as an important predictor for participation in Wallop, a result that is not surprising given the inherently social nature of the Wallop service.

The most surprising results from the first two models are the impact of mutual ties, which are correlated with decreased activity regardless of the relative activity level of the other user involved in the connection. Mutual connections to more active users should be predictive of a subsequent increase in participation, while mutual ties to less active users should be predictive of a subsequent decrease in participation. Yet there is a strong and consistent negative relationship between mutual ties to more active users and subsequent participation rates. Adding controls for tie strength does not help answer this riddle, as the impact of mutual tie strength with more active users is not significant, and the effect of the number of such mutual ties is magnified. On the other hand, increasing the number of mutual connections with less active users is predictive of a decrease in participation, but when tie strength is included in the model the number of mutual connections loses significance.

One possible explanation for the results of mutual ties is reciprocation. Most ties in the Wallop comment network are reciprocated, suggesting the possibility of a norm of reciprocity. If such a norm exists, a user receiving messages during a given

time period may log in to Wallop more frequently than usual in order to adhere to the norm of reciprocity. The subsequent return to more normal levels of activity might account for some of the negative impact of this measure.

By contrast, models 3-5 suggest a general correlation between interaction with closed triads and increased participation in Wallop. In particular, the impact of triadic closure amongst mutual ties is consistently positive, although only the effect of triadic closure amongst less active neighbors is robust and significant. This positive signal suggests social groups are important for encouraging participation in Wallop. As with Wikipedia, the presence of a community of active participants may provide some affirmation for a user that time spent on Wallop is time well spent. These signals are more consistent in Wallop than the effects observed in the Wikipedia analysis. This might be a result of the more socially oriented nature of the Wallop service, as the increased proportion of closed triads in Wallop than Wikipedia could result in more robust measures.

Including controls for tie strength affects the importance of triadic closure, causing a small decrease in the importance of triadic closure where both connections are mutual ties to less active users. The overall effect of triadic closure remains positive, significant, and meaningful. Triadic closure, while capturing some of the same effects as tie strength, has an impact in its own right in the Wallop data, a result which provides additional support for the importance of social groups and local clustering as a predictor of participation.

The impact of the control variables in this analysis is similar to the impact of the control variables in the Wikipedia analysis. Wallop users acting during the first three months of the observation window are more likely to increase contribution, probably due to novelty effects and the overall growth in the Wallop user base. However, by the end of the observation window Wallop itself was clearly a dying

system, and that is reflected in the strong negative coefficient associated with activity occurring in the final time period.

Generally speaking, long-term users are more likely to decrease participation, although this effect might be driven by heavily active users exiting the system abruptly as the service declined towards the end of the observation window. Activity in the previous time period is also negatively correlated with subsequent change in participation, although this is at least partly an artifact of the way the metric is defined. Since activity is bounded by the number of days in the time period, the odds of increasing activity actually go down as the number of days active goes up.

Summary of Findings

These findings corroborate some of the results from the Wikipedia study. Social relationships are important as predictors of subsequent participation. While degree measures do not necessarily have the expected effect, measures of tie strength and local structure illustrate the importance of social connections for participation. Triadic closure among mutual connections is consistently predictive of increased participation, suggesting the importance of a social group. If a user is connected to a social group within the Wallop system, s/he will generally increase participation even if that group is itself less active than s/he is.

The Wikipedia analysis shows that social ties are important given certain local network properties, and although the effects are not particularly strong it appears social groups are important for subsequent participation. In a task-oriented setting like Wikipedia, the effect of clustering in a local network may trump relative activity levels. The results of the Wallop study confirm the importance of the social group. Indeed, in a socially oriented system like Wallop the effects of triadic closure and local network structure are more consistent than in Wikipedia. Furthermore, in both

systems the effects of triadic closure remain important even as controls for tie strength are included in the models, suggesting that triadic closure is not a simple proxy for tie strength but rather captures an important feature of a user's social network. Further analysis of a variety of task-oriented and socially oriented groups would be instructive, especially if the relative costs of contribution could be varied in a controlled fashion.

CHAPTER 5:
CONCLUSION

Summary

The study presented here examines four basic questions. Do social relationships predict continued participation in a low-cost voluntary association? How is triadic closure related to an individual's decision to participate? If there is a triadic closure effect, is it simply encoding dyadic tie strength or is it an important predictor in its own right? Do these effects depend on whether an individual is participating in a task-oriented or socially oriented setting?

The answers are presented in an analysis of two distinct data sets. The first, Wikipedia, is a task-oriented community-created encyclopedia. Although individuals socialize with one another within the Wikipedia system, the primary goal of the group is to produce a useful encyclopedia. The second, Wallop, is a socially oriented blogging and social networking service. In Wallop, the group is the goal. Individuals participate in order to socialize with other users and share photos, comments, thoughts, and ideas with one another.

The analysis of each system was conducted by deriving the underlying social network structure from directed communication taking place within each system. This network was used to generate metrics for dyadic relationships, tie strength, and triadic closure. These variables were then used as key predictors in a longitudinal model, and their relative predictive power was extracted from the results.

The remainder of this chapter presents a discussion of the findings from this analysis. It then describes the key limitations of the research in a bit more detail, and presents suggestions for future research to address these limitations. It concludes with a brief summary of the research.

Discussion of Results

Overall, there is mixed support for the notion of dyadic social ties as predictors of continued participation. However, triadic closure is generally predictive of increasing participation, even in cases where the connected neighbors are less active than the focal user. This suggests a possible social affirmation effect, where individuals are encouraged to participate by connections embedded in a social group, regardless of the level of activity of the group members. These effects are not eliminated by the inclusion of tie strength measures, suggesting that triadic closure is encoding some effect independent from tie strength. Finally, the primary difference between Wallop and Wikipedia appears to be the effect of tie strength. In Wikipedia, increased tie strength is consistently predictive of decreased participation, suggesting the possibility that increased social activity distracts participants from making contributions within a task-oriented system.

Impact of Social Ties

The overall results suggest mixed support for the theoretical expectations around dyadic ties as predictors of continued participation. As expected, mutual connections to less active users are negatively associated with participation. However, in both the Wikipedia and Wallop data this effect is weak and not robustly significant, and the effects for mutual connections to more active neighbors do not match expectations.

In Wikipedia, mutual connections with more active users have the predicted positive relationship with subsequent participation. However, increased tie strength is predictive of a subsequent decrease in participation. One possible reason for this effect is that the stronger social relationships are distracting the user from contributing to the

encyclopedia. It is also possible that stronger ties indicate an antagonistic social relationship, which could encourage a user to stop contributing altogether.

The converse is true in Wallop. Counter to expectations, as the number of mutual connections with more active users increases, an individual's subsequent activity decreases. This effect may be related to a norm of reciprocation. In Wallop, messages are normally reciprocated, and therefore mutual ties to more active users may indicate simple reciprocation on the part of a user who rarely contributes content. For such users the act of sending the message will increase their current activity rate, which could lead to a subsequent decrease as their activity returns to normal.

One-way connections do not necessarily show the expected results, but they are also likely to be unreliable indicators as they may be encoding some other factor affecting individual participation. For example, unreciprocated incoming ties may indicate that the target user has already decided to stop using the service. This explanation is supported by the consistent negative signal associated with incoming ties in both Wallop and Wikipedia. On the other hand, unreciprocated outward ties might indicate a heavily active and invested user, or they might indicate a set of ties awaiting an incoming message in the subsequent time period. In the former case, a negative relationship with subsequent interaction may simply be a result of the user being too active to reasonably increase activity. In the latter case, a positive impact on individual activity may be a result of the user responding to the reciprocated message.

Although the findings are a bit mixed, there is some support for the notion that dyadic connections to active participants are predictive of subsequent increases in participation. In a socially oriented setting, increasing dyadic tie strength to active participants predicts a small increase in activity, although increasing the raw number of ties has the opposite effect. In a task-oriented setting increasing tie strength is negatively associated with subsequent participation. This may be the result of social

activity cannibalizing the time an individual would have spent contributing to the task, a problem that does not exist in a setting where the primary goal of the group is to engage in social activity.

Importance of Local Structure

In general, increased levels of triadic closure predict a subsequent increase in participation rate. Contrary to expectations, this is true whether the triads are made up of more active or less active alters. This suggests that triadic closure indicates some form of social reinforcement. An individual with relationships embedded within some social group might well be relying on that group to affirm her participation as worthwhile or valuable in some way. Even though the users she is connected to are less active, she may well feel that her time is well spent as long as she is connected to a cohesive social group engaged in the same activity.

Including controls for tie strength within the triads diminished the triadic closure effect somewhat, although it did not come close to eradicating it. Therefore, tie strength and triadic closure are not encoding the same effect, although there may be some degree of overlap between the two metrics. The triadic tie strength measures themselves have no significant effect in the Wallop data, but do have some negative effect in the Wikipedia data. This once again suggests the possibility that social interaction distracts participants from contributing to the task, as users who get more deeply embedded in the social aspects of Wikipedia decrease their contributions to the encyclopedia section of the site.

Task Oriented and Socially Oriented Systems

Overall, the results for Wallop and Wikipedia are fairly similar. In both cases, connections to a cohesive social group are important regardless of relative rates of

activity. This suggests that both task oriented and socially oriented settings can benefit from a certain degree of social reinforcement, and that establishing a well connected social group within a voluntary association can be beneficial.

The key differences between Wallop and Wikipedia are related to the relative importance of tie strength. In Wallop tie strength is generally either unimportant or has the expected effects. In Wikipedia strong ties and strong triads are negatively associated with subsequent contribution. This suggests that being part of the social fabric of a task-oriented group is beneficial for contribution, perhaps due to some affirmation of the value of contributing. However, when an individual spends too much time socializing it may become a distraction and actually reduce contribution to the task, a problem that clearly does not exist in groups where social interaction is the primary goal.

Limitations and Future Work

There are five key limitations to this study. Some of these limitations cannot be addressed without additional data, but others could potentially be addressed with a follow-up study using the same data sets. Such follow-up studies are beyond the scope of the research presented here, but based on the results of this study they are likely worthwhile extensions to pursue.

The first key limitation of this work is the lack of demographic data. Demographic factors, such as age, gender, and socio-economic status might influence individual participation in Wikipedia or Wallop in much the same way they influence user participation in other online services (see e.g., Fallows (2005)). Although the fixed-effects model mitigates a lot of the first-order concern regarding demographic differences between individuals, there is still some potential for a life-changing event to alter an individual's participation rate. Furthermore, there are second-order

considerations as well. It is entirely possible that some edges are more important for a given individual than others. For example, if there is a high degree of homophily between the focal actor and one of the other users she communicates with, then that edge may be more relevant to the individual and therefore more predictive of future participation than the other edges. Indeed, triadic closure may be encapsulating the demographic similarity between two connections more than the raw strength of those ties as measured by volume of messages.

This issue of homophily across an edge is an example of the second important limitation of this research. Although the tie strength metric used in this analysis is reasonable, there are other possible metrics that might also be conflated with triadic closure. One suggestion for future study is to find a data set with more detailed demographic data and replicate this study with demographic controls included. In addition to including demographic controls, testing other measures of tie strength – both those based on demographic attributes and other network attributes – to determine whether or not the effects of triadic closure remain robust to other tie strength metrics.

Another weakness of this research is the lack of rigorous content analysis. Although the high level statistical analysis presented here is illuminating, it does miss certain important details. For example, ties may have valence, but without content analysis there is no way to determine what the users are saying in their interactions with one another. If a tie is strong and negative it could have the opposite effect of a strong positive tie. Furthermore, the distinction between the task section of Wikipedia and the social section of Wikipedia may not be as clear as it seems. A certain amount of task-related discussion takes place in the user talk sections of Wikipedia, and it is entirely possible that a certain amount of social interaction takes place in the article

talk sections. Extending this study to include a content analysis component could resolve both of these concerns, or at least allow for a more tightly controlled analysis.

Perhaps the most problematic issue from the perspective of comparing socially oriented and task oriented systems is the fact that this study conflates social interaction with participation in the Wallop data, but does not do the same thing in the Wikipedia data. The reason for this inconsistency is the clear separation between the task and social components in the Wikipedia interface, and the inherent overlap between social activity and participation in Wallop. While there may well be some overlap between the two areas, the sharp distinction in Wikipedia combined with the fact that contributions generally dwarf social interactions make separation a logical choice. Unfortunately, Wallop includes no such separation – social interaction and participation are inextricably linked, in part because a reasonably large percentage of content contribution events were in fact part of the social interaction component of the service. The primary finding – that triadic closure is important regardless of relative activity levels – should not be meaningfully impacted by this inconsistency, but it may cloud the interpretation of how task and socially oriented systems compare with one another. An additional study using a socially oriented system that allows for a clear separation between the social interaction activities and more general participation would be ideal, as it would allow for a more direct, more systematic comparison of the effects studied here.

Finally, it is not clear that the results obtained here generalize well to offline settings. Although the triadic closure findings are consistent with existing research on LiveJournal, it is possible that this is an artifact of the online medium. The fact that other researchers have found similar effects in telephone communications data (Eagle et al., 2010) makes this less likely, but additional study in offline settings is warranted.

This analysis could also explore a range of environments, with varying costs to contribution and varying degrees of task orientation.

Conclusion

This research focuses on questions surrounding the relationship between social ties and participation in a voluntary association. Early findings in the social movement literature (McAdam, 1986; Snow et al., 1980) find strong dyadic relationships to be important predictors of participation. Later results highlight the importance of triadic closure (Backstrom et al., 2006), a finding that might also explain the results obtained in earlier studies where triadic closure was the basic measure of tie strength (McAdam, 1986).

The study presented here first attempts to answer the open question of whether or not the triadic closure effects observed by Backstrom et al. (2006)are a result of triadic closure encoding tie strength. The data reinforces earlier findings regarding the importance of triadic closure as a predictor of subsequent participation. Furthermore, triadic closure and tie strength appear to be encoding different effects, as the effects of triadic closure remain important when controls for tie strength are included in the model. Additional research should address the robustness of the results and attempt to resolve some of the limitations presented here.

REFERENCES

1. Adamic, L. A. & Adar, E. (2003) Friends and Neighbors on the Web. *Social Networks*. 25, 211-230.

2. Adamic, L. A. & Glance, N. (2005) The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. *WWW2005, 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis, and Dynamics*. 16 Pages.

3. Adar, E. & Adamic, L. A. (2005) Tracking Information Epidemics in Blogspace. *Web Intelligence 2005*.

4. Anderson, R. & May, R. (1992) Infections Diseases of Humans - Dyanmics and Control. Oxford University Press.

5. Anheier, H. (2003) Movement Development and Organizational Networks: The Role of 'Single Members' in the German Nazi Party, 1925-30. *Social Movements and Networks: Relational Approaches to Collective Action*. Pages 49-76. McAdam, Douglas and Diani, Mario, eds. Oxford University Press.

6. Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006) Group Formation in Large Social Networks: Membership, Growth, and Evolution. *KDD '06*.

7. Becker, P. E. (1999) *Congregations in Conflict: Cultural Models of Local Religious Life*. Cambridge University Press.

8. Bimber, B. (2000) Measuring the Gender Gap on the Internet. *Social Science Quarterly*. 81, 868-876.

9. Blood, R. (2004) How Blogging Software Reshapes the Online Community. *Communications of the ACM*. 47, 53-55.

10. Bott, E. (1957) *Family and Social Network*. Routledge.

11. Brown, R. K. & Brown, R. E. (2003) Faith and Works: Church-Based Social Capital Resources and African-American Political Activism. *Social Forces*. 82, 617-641.

12. Bryant, S. L., Forte, A., & Bruckman, A. (2005) Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. *ACM SIGGROUP Conference on Supporting Group Work*.

13. Burke, M., Joyce, E., Kim, T., Anand, V., & Kraut, R. E. (2007) Introductions and Requests: Rhetorical Strategies that Elicit Community Response. *Communities and Technologies 2007*.

14. Burke, M., Marlow, C., & Lento, T. M. (2009) Feed Me: Motivating Newcomer Contribution in Social Network Sites. *ACM CHI 2009: Conference on Human Factors in Computing Systems*.

15. Burke, M., Marlow, C., & Lento, T. M. (2010) Social Network Activity and Social Well-Being. *ACM CHI 2010: Conference on Human Factors in Computing Systems*.

16. Burt, R. S. (1987) Social Contagion and Innovation: Cohesion Versus Structural Equivalence. *American Journal of Sociology*. 92, 1287-1335.

17. Cantril, H. (1941) *The Psychology of Social Movements.* Wiley and Sons.

18. Caverlee, J. & Webb, S. (2008) A Large-Scale Study of MySpace: Observations and Implications for Online Social Networks. *ICWSM 2008*.

19. Centola, D. & Macy, M. W. (2007) Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology*. 113, 702-34.

20. Christakis, N. A. & Fowler, J. H. (2007) The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine*. 357, 370-9.

21. Chwe, M. S.-Y. (1999) Structure and Strategy in Collective Action. *American Journal of Sociology*. 105, 128-56.

22. Coleman, J. (1988) Social Capital in the Creation of Human Capital. *American Journal of Sociology*. 94, S95-S120.

23. Coleman, J., Katz, E., & Menzel, H. (1957) The Diffusion of an Innovation Among Physicians. *Sociometry*. 20, 253-270.

24. Cummings, J. N., Butler, B., & Kraut, R. E. (2002) The quality of online social relationships. *Communications of the ACM*. 45, 103-108.

25. Des Jarlais, D. C., Arasteh, K., Perlis, T., Hagan, H., Heckathorn, D., McKnight, C., Bramson, H., & Friedman, S. R. (2007) The transition from injection to non-injection drug use: long-term outcomes among heroin and cocaine users in New York City. *Addiction*. 102, 778-85.

26. DiMaggio, P. J., Hargittai, E., Neuman, W. R., & Robinson, J. (2001) Social Implications of the Internet. *Annual Review of Sociology*. 27, 307-336.

27. Dimmick, J., Kline, S., & Stafford, L. (2000) The gratification niches of personal e-mail and the telephone: Competition, displacement and complementarity. *Communication Research*. 27, 227-248.

28. Dixon, M. & Roscigno, V. (2003) Status, Networks and Social Movement Participation: The Case of Striking Workers. *American Journal of Sociology*. 108, 1292-1327.

29. Dobransky, K. & Hargittai, E. (2006) The Disability Divide in Internet Access and Use. *Information, Communication and Society*. 9, 313-334.

30. Eagle, N., Macy, M., & Claxton, R. (2010) Network Diversity and Economic Development. Unpublished Manuscript.

31. Ellison, N., Steinfeld, C., & Lampe, C. (2007) The benefits of Facebook 'friends': Exploring the relationship between college students' use of online social networks and social capital. *Journal of Computer-Mediated Communication*. Volume 12.

32. Fallows, D. (2005) How Women and Men Use the Internet. *Pew Internet and American Life Project*.

33. Feinberg, W. E. & Johnson, N. R. (1988) Outside Agitators and Crowds: Results from a Computer Simulation Model. *Social Forces*. 67, 398-423.

34. Fernandez, R. M. & McAdam, D. (1988) Social Networks and Social Movements: Multiorganizational Fields and Recruitment to Mississippi Freedom Summer. *Sociological Forum*. 3, 357-382.

35. Fisher, D., Smith, M., & Welser, H. T. (2006) You are who you talk to: Detecting roles in usenet groups. *Proceedings of the 39th Hawaii International Conference on Systems Sciences (HICSS)*.

36. Goel, S., Muhamad, R., & Watts, D. (2009) Social Search in "Small-World" Experiments. *WWW 2009*.

37. Goel, S. & Salganik, M. J. (2010) Assessing Respondent-Driven Sampling. *Proceedings of the National Academy of Sciences*. 107, 6743-6747.

38. Gould, R. V. (1993) Collective Action and Network Structure. *American Sociological Review*. 58, 182-197.

39. Gould, R. V. (1991) Multiple Networks and Mobilization in the Paris Commune, 1871. *American Sociological Review*. 56, 716-729.

40. Granovetter, M. (1973) The Strength of Weak Ties. *American Journal of Sociology*. 78, 1360-1380.

41. Granovetter, M. (1978) Threshold Models of Collective Behavior. *American Journal of Sociology*. 83, 1420-1443.

42. Gu, L., Johns, P., Lento, T. M., & Smith, M. A. (2006) How do Blog Gardens Grow? Language community correlates with network diffusion and adoption of blogging systems. *Proceedings of the AAAI Symposium on Computational Approaches to Analyzing Weblogs*.

43. Hargittai, E. (2008) The Digital Reproduction of Inequality. *Social Stratification*. Pages 936-944. Grusky, David ed. Westview Press.

44. Heckathorn, D. (1997) Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*. 44, 174-99.

45. Heckathorn, D. (2002) Respondent-driven sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. *Social Problems*. 49, 11.

46. Hedstrom, P., Sandell, R., & Stern, C. (2000) Mesolevel Networks and the Diffusion of Social Movements: The Case of the Swedish Social Democratic Party. *American Journal of Sociology*. 106, 145-72.

47. Herring, S., Job-Sluder, K., Scheckler, R., & Barab, S. (2002) Searching for Safety Online: Managing "Trolling" in a Feminist Forum. *The Information Society*. 18, 371-384.

48. Herring, S., Kouper, I., Paolillo, J. C., Scheidt, L. A., Tyworth, M., Welsch, P., Wright, E., & Yu, N. (2005) Conversations in the Blogosphere: An Analysis "From the Bottom Up". *Hawai'i International Conference on Systems Sciences*.

49. Herring, S., Scheidt, L. A., Bonus, S., & Wright, E. (2004) Bridging the Gap: A Genre Analysis of Weblogs. *Hawai'i International Conference on Systems Science*.

50. Johnson, N. R. & Feinberg, W. E. (1977) A Computer Simulation of the Emergence of Consensus in Crowds. *American Sociological Review*. 42, 505-521.

51. Kale, A., Karandikar, A., Kolari, P., Java, A., Finin, T., & Joshi, A. (2007) Modeling Trust and Influence in the Blogosphere Using Link Polarity. *International Conference on Weblogging and Social Media (ICWSM 2007).*

52. Kim, H. & Bearman, P. S. (1997) The Structure and Dynamics of Movement Participation. *American Sociological Review*. 62, 70-93.

53. Kolari, P., Finin, T., & Joshi, A. (2006) SVMs for the Blogosphere: Blog Identification and Splog Detection. *Proceedings of the AAAI Symposium on Computational Approaches to Analyzing Weblogs*.

54. Kollock, P. & Smith, M. (1996) Managing the Virtual Commons: Cooperation and Conflict in Computer Communities. *Computer-Mediated Communication.* Herring, Susan, ed. John Benjamins.

55. Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2003) On the bursty evolution of blogspace. *WWW2003*.

56. Kumar, R., Novak, J., & Tomkins, A. (2006) Structure and Evolution of Online Social Networks. *KDD '06*.

57. Lai, G. & Wong, O. (2002) The Tie Effect on Information Dissemination: The Spread of a Commercial Rumor in Hong Kong. *Social Networks*. 24, 49-75.

58. Lampe, C. & Johnston, E. (2005) Follow the (Slash) dot: Effects of Feedback on New Members in an Online Community. *International Conference on Supporting Group Work (GROUP '05).*

59. Lento, T. M., Welser, H. T., Gu, L., & Smith, M. A. (2006) The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop Weblogging System. *WWW 2006: 3rd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis, and Dynamics*.

60. Leskovec, J., Adamic, L. A., & Huberman, B. A. (2007) The Dynamics of Viral Marketing. *7th ACM Conference on Electronic Commerce.*

61. Leskovec, J., Singh, A., & Kleinberg, J. (2006) Patterns of Influence in a Recommendation Network. *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD).*

62. Macy, M. (1991) Chains of Cooperation: Threshold Effects in Collective Action. *American Sociological Review*. 56, 730-747.

63. Marsden, P. V. (1990) Network Data and Measurement. *Annual Review of Sociology*. 16, 435.

64. Marsden, P. V. & Campbell, K. (1984) Measuring Tie Strength. *Social Forces*. 63, 482-501.

65. McAdam, D. (1986) Recruitment to High-Risk Activisim: The Case of Freedom Summer. *American Journal of Sociology*. 92, 64-90.

66. McAdam, D. (1988) *Freedom Summer*. Oxford University Press.

67. McAdam, D. & Paulsen, R. (1993) Specifying the Relationship between Social Ties and Activism. *American Journal of Sociology*. 99, 640-667.

68. Moody, J. (2002) The Importance of Relationship Timing for Diffusion. *Social Forces*. 81, 25.

69. Myers, D. J. (2000) The diffusion of collective violence: infectiousness, susceptibility, and mass media networks. *American Journal of Sociology*. 106, 173-208.

70. Nardi, B. A., Schiano, D. J., Gumbrecht, M., & Swartz, L. (2004) Why We Blog. *Communications of the ACM*. 47, 41-49.

71. Newman, M. E. J. (2002) Spread of Epidemic Disease on Networks. *Phys. Rev. E*. 66, 11.

72. Oliver, P. & Marwell, G. (2001) Whatever Happened to Critical Mass Theory? A Retrospective and Assessment. *Sociological Theory*. 19, 292-311.

73. Oliver, P. & Marwell, G. (1988) The Paradox of Group Size in Collective Action: A Theory of the Critical Mass. II.. *American Sociological Review*. 53, 1-8.

74. Oliver, P., Marwell, G., & Prahl, R. (1988) Social Networks and Collective Action: A Theory of the Critical Mass. III. *American Journal of Sociology*. 94, 502-34.

75. Oliver, P., Marwell, G., & Teixeira, R. (1985) A Theory of the Critical Mass. I. Interdependence, Group Heterogeneity, and the Production of Collective Action. *American Journal of Sociology*. 91, 522-56.

76. Oliver, P. E. (1993) Formal Models of Collective Action. *Annual Review of Sociology*. 19, 271-300.

77. Oliver, P. E. & Myers, D. J. (2003) Networks, Diffusion, and Cycles of Collective Action. *Social Movements and Networks: Relational Approaches to Collective Action*. Pages 173-203. McAdam, Douglas and Diani, Mario, eds. Oxford University Press.

78. Olson, M. (1965) *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press.

79. Opp, K.-D. & Gern, C. (1993) Dissident Groups, Personal Networks, and Spontaneous Cooperation: The East German Revolution of 1989. *American Sociological Review*. 58, 659-680.

80. Opp, K.-D. & Roehl, W. (1990) Repression, Micromobilization, and Political Protest. *Social Forces*. 69, 521-547.

81. Postmes, T., Spears, R., & Lea, M. (2000) The Formation of Group Norms in Computer-Mediated Communication. *Human Communication Research*. 26, 341-371.

82. Preece, J. & Maloney-Krichmar, D. (2003) Online Communities. *Handbook of Human-Computer Interaction*. Pages 596-620. Jacko, J and Sears, A, eds. Lawrence Erlbaum Associates, Inc.

83. Ramirez-Valles, J., Garcia, D., Campbell, R. T., Diaz, R. M., & Heckathorn, D. (2008) HIV Infection, Sexual Risk Behavior, and Substance Use Among Latino Gay and Bisexual Men and Transgender Persons. *American Journal of Public Health*. 98, 1036-1042.

84. Salganik, M. J., Dodds, P. S., & Watts, D. (2006) Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*. 311, 854-856.

85. Salganik, M. J. & Watts, D. (2009) Web-Based Experiments for the Study of Collective Social Dynamics in Cultural Markets. *Topics in Cognitive Science*. 1, 439-468.

86. Schiano, D. J., Nardi, B. A., Gumbrecht, M., & Swartz, L. (2004) Blogging by the rest of us. *Ext. Abstracts of CHI '04*. 1143-1146.

87. Skvoretz, J. & Fararo, T. (1996) Status and Participation in Task Groups: A Dynamic Network Model. *American Journal of Sociology*. 101, 1366-1414.

88. Snow, D. A., Zurcher, L. A., & Ekland-Olson, S. (1980) Social Networks and Social Movements: A Microstructural Approach to Differential Recruitment. *American Sociological Review*. 45, 787-801.

89. Steinfeld, C., DiMicco, J. M., Ellison, N., & Lampe, C. (2009) Bowling Online: Social Networking and Social Capital within the Organization. *Proceedings of the Fourth Communities and Technologies Conference*.

90. Steinfeld, C., Ellison, N., & Lampe, C. (2008) Social Capital, Self-Esteem, and Use of Online Social Network Sites: A Longitudinal Analysis. *Journal of Applied Developmental Psychology*. 29.

91. Strang, D. & Macy, M. (2001) In Search of Excellence: Fads, Success Stories, and Adaptive Emulation. *American Journal of Sociology*. 107, 147-82.

92. Strang, D. & Tuma, N. B. (1993) Spatial and Temporal Heterogeneity in Diffusion. *American Journal of Sociology*. 99, 614-639.

93. Toch, H. (1965) The Social Psychology of Social Movements. Bobbs-Merrill.

94. Valente, T. W. (1996) Social Network Thresholds in the Diffusion of Innovations. *Social Networks*. 18, 69-89.

95. Van den Bulte, C. & Lilien, G. L. (2001) Medical Innovation Revisited: Social Contagion versus Marketing Effort. *American Journal of Sociology*. 106, 1409-35.

96. Viegas, F., Wattenberg, M., Kriss, J., & van Ham, F. (2007) Talk Before You Type: Coordination in Wikipedia. *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*.

97. Watts, D. (2002) A Simple Model of Global Cascades on Random Networks. *Proceedings of the National Academy of Sciences*. 99, 5766-5771.

98. Watts, D. (1999) Network dynamics and the small world phenomenon. *American Journal of Sociology*. 105, 493-527.

99. Watts, D. (2004) The "New" Science of Networks. *Annual Review of Sociology*. 30, 243-70.

100. Watts, D. & Strogatz, S. H. (1998) Collective Dynamics of 'Small-World' Networks. *Nature*. 393, 440-442.

101. Wellman, B. (1996) An Electronic Group is Virtually a Social Network. *Culture of the Internet*. Pages 179-208. Kiesler, Sara, ed. Lawrence Erlbaum.

102. Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., & Haythornthwaite, C. (1996) Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community. *Annual Review of Sociology*. 22, 213-38.

103. Welser, H. T., Gleave, E., Fisher, D., & Smith, M. (2007) Visualizing the Signatures of Social Roles in Online Discussion Groups. *The Journal of Social Structure*. 8, 2.

104. Wrigley, N. (1990) Unobserved Heterogeneity and the Analysis of Longitudinal Spatial Choice Data. *European Journal of Population*. 6, 327-358.

105. Zachary, W. W. (1977) An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*. 33, 452-473.

106. Zohoori, N. & Savitz, D. A. (1997) Econometric Approaches to Epidemiologic Data: Relating Endogeneity and Unobserved Heterogeneity to Confounding. *Annals of Epidemiology*. 7, 251-257.