Formation of Citation and Reply ties over intra-organizational blog network[1]

Nachiketa Sahoo, Ramayya Krishnan

Heinz School

Carnegie Mellon University

Jamie Callan

Language Technologies Institute

Carnegie Mellon University

Abstract

The effect of brick-and-mortar and online factors on the formation of citation and reply ties is studied using an unique data set that compiles blogging activity within a global IT services firm for over a year and a half. We found that online ties over the blog network are not limited by the brick-and-mortar ties. Rather, they behave as a cohesive, persistent, and independent group. Further examination of online ties reveals interesting interaction among different types of ties. We observe the transformation of reply-together ties into direct reply ties over time. This illustrates the effect of lateral communication in the blogs. We observe a very interesting negative effect of citations ties on future reply ties suggesting that certain pairs of actors might be substituting one type of tie for the other. We also find positive effect of the reciprocal and simmelian conditions as predicted by prior work in social sciences. However, having multi-clique simmelian ties between a pair of individuals has a negative effect on their future interaction. The insights from these findings are discussed in the light of the characteristic of different ties.

[1] This is a student authored paper.

1 Introduction

Increasingly organizations are creating private blogs to effectively engage with their employees to capture knowledge and disseminate information in a timely manner (Ojala 2005). However, they run the risk of only managing a segregated blog network where employees interact only with those they already know through their brick-and-mortar network—after all, studies have shown that Homophily has strong influence on a person's social networks (McPherson, Smith-Lovin et al. 2001). In such a case the information flow over the blog network will be restricted and the benefit to the organization will be minimal. Therefore, it is important to know the extent to which the brick-and-mortar network influences the online network. In the first part of this study we examine this with the help of an unique longitudinal intra-organizational blog dataset collected from a global IT Software Services firm spanning over a year and half (Jul '06—Jan '08).

In the second part we examine the interaction between the online ties over the blog network. The blog social network is a collection of multiple interacting ties. It can be argued that the most common tie formed between bloggers is a *reading tie*. Reading blogs is a habit of many with each following a favorite set of blogs regularly. As common as it may be, it is one of the hardest to measure and we do not observe reading behavior in our data set. However, reading potentially gives rise to two important ties that are readily observed, namely, reply ties and citation ties. Blog posts are discussed and sustained by the replies from the readers. The nature of a reply tie is conversational. On the other hand, citations give credence to a blog post and spread information over the network. Thus, repeated citations from one person to another suggest that the person citing considers the posts of the other to be noteworthy. Understanding what drives or stifles the

formation of these two building blocks of the blog social network will enable us to effectively implement a blog network inside an organization.

1.1 Background

Link formation in social networks has been studied before in Data Mining literature. Nowell and Kleinberg has explored the proximity measures that best predict the future online link formation using a co-authorship network over academic publications (Liben-Nowell and Kleinberg 2007). McGlohon et al has proposed methods to cluster public blogs based on their posting and linking behavior (McGlohon, Leskovec et al. 2007). Getoor and Diehl has surveyed the data mining literature for node ranking and classification, group detection, link prediction, subgraph discovery among other common network mining tasks (Lise and Christopher 2005). In social sciences social network analysis methods have been very well developed (Wasserman and Faust 1994). Although, its methods such as Multliple Regression Quadratic Assignment Procedure(Krackhardt 1987), Double Dekker Semi Partialing (Dekker, Krackhardt et al. 2007) among many others, have been primarily developed in the context of brick-and-mortar networks, hence do not scale beyond a few thousand nodes[2], they can reveal complex interactions among the nodes and ties in a social network.

1.2 Contribution

This paper makes a contribution at the intersection the of weblog mining and social network analysis literature. We analyze the corporate weblogs using social network analysis methods to:

1. Compare the effect of brick-and-mortar ties and online ties on the formation of future citation and reply ties. This sheds light on the extent to which the online networks are

[2] In fact, the current study involving over three thousand nodes and up to eight edge variables is one of the largest we have seen where such complex methods have been successfully applied.

limited by traditional brick-and-mortar networks. It will also help us identify the most important factors affecting the citation and reply ties.

2. Study the effect of different online ties and the tie conditions such as reciprocated ties, Simmelian ties (Simmel 1950), and co-participation on future citation and reply ties. This will help us understand the dynamic interaction of various online ties.

2 Data preparation

The data for this study is collected from an employee-only blog network in a large IT services firm between Jul '06 and Jan '08. It contains the blog posts and replies along with timestamps and employment information of the bloggers (Summarized in Table 1).

Date range	Jul '06—Jan '08	
Blog posts	51,000, average length 1778 characters	
Comments	131,000, average length 170 characters	
Bloggers	3700	
Demographic Variables	Employee Id Location Experience Designation Supervisor	 359 unique 1-29 years 144 unique
Blogs	2950	

Table 1 Data description

We prepare the data for a *two period* study so that we can measure the effect of relations observed in the earlier period on formation of citation and reply ties in the later. As we can see from Figure 1 most of the blog posts and replies have occurred in the last one year. We split the dataset into two approximately equal halves in terms of the volume of blog posts and replies. One half contains posts and replies incident before 1st Oct 2007 and the other contains those that occurred after.

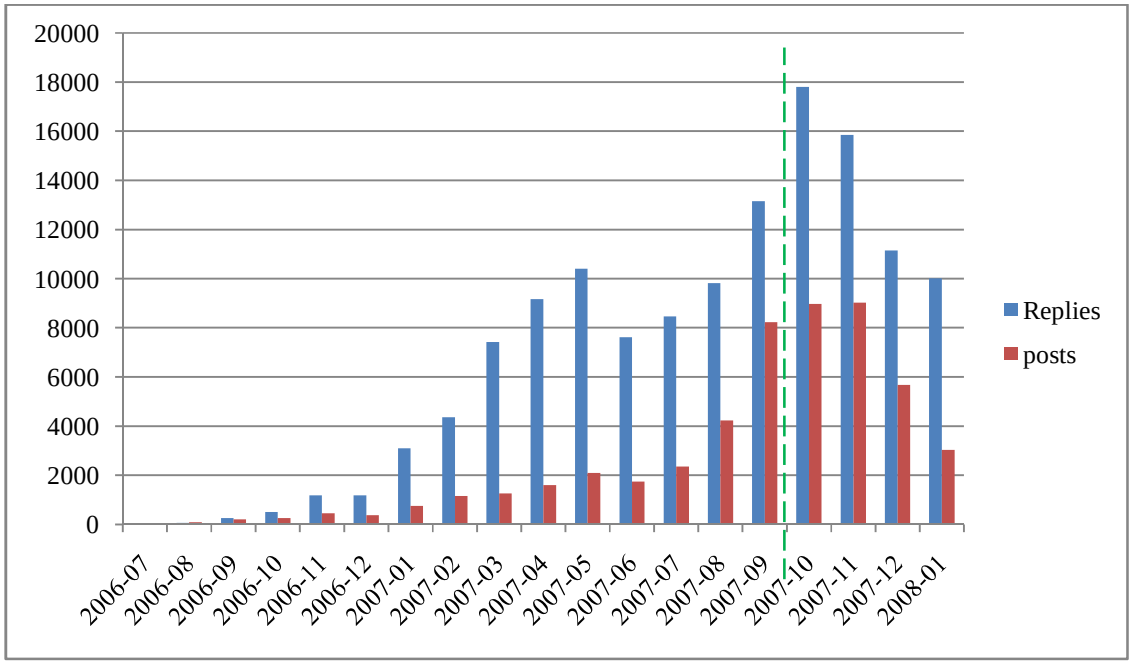

Figure 1 Blogging activity over 19 months

3 Theoretical framework

3.1 Types of Ties and Networks

There are several *types* of ties among the employee bloggers. This leads to several overlaying networks among them. Some of them are:

1. Reporting network derivatives

    a. *Who reports to whom:* It is a directed network (Figure 4). Reporting ties are one of the strongest ties in the organization. They enforce accountability for responsibilities within the hierarchy and play an important role in the organization achieving its goals. Therefore, two people connected by a direct reporting tie are closely connected.

b. *Reporting sibling* or *same boss*: It is an undirected network derived from the previous network. Often times a group of people work together in a project reporting to the same boss. Thus, when two employees report to the same boss they are likely to have a strong work related tie.

c. *Reporting roles network:* It is a directed network of roles reporting to other roles (Figure 5). In addition of showing the approximate level of different designations, the in-degree and the out-degree of the nodes tell us how specialized the role is.

2. Organization affiliation networks

These undirected networks are formed between employees through their common affiliations in the organization.

a. *Co-location:* Employees going to same offices usually work in a common functional division of the organization. They tend to have access to similar information channels. In an organization with geographically distributed offices, different offices often have different culture. Thus two employees going to the same office have a tie by the virtue of their work location.

b. *Common rank*: Employees having the same rank within the organizations are likely to interact more than those with widely differing ranks.

3. Reply network

a. *Who replies to whom:* Bloggers are seen to have a persisting reply pattern—usually each chooses to participate at only a few blogs. Thus, the reply activity

leads to a directed tie between two employees. Strength of this relation can be measured by the number of replies or number of replies per day.

b. *Who reply together:* Blogs are used not only as a bidirectional channel of communication between the blogger and the others who comment, but also as a channel for those who comment to discuss among themselves. Hence, if two employees are seen to reply together at different blogs, then it is likely that they will come to know each other and possibly develop other types of ties in the future.

4. Citation network (*Who cites who*)

   Bloggers cite other bloggers in their posts through hyperlinks. A citation can be seen as an endorsement or as a grant of credence to post cited. We find that this is a much rarer activity than the reply activity.

Some of these are direct ties, e.g., reporting, reply, citation etc. However, others are derived ties, e.g., pairs that reply together, pairs that report to the same boss or go to the same office.

3.2 Reciprocal ties and Simmelian ties

Two tie configurations merit further examination. When the ties are reciprocated they signal a two way participation leading to much stronger relation between the two actors than when the ties are directed and one sided. However, when such reciprocated ties are part of a clique they exhibit yet different properties, because, in such a configuration the ties begin to be governed by group norms (Simmel 1950). Simmel notes that this change is important because, adding one node to a dyad fundamentally changes the nature of the tie, whereas, growing this clique does not change the tie further by as much. Such ties are known as Simmelian ties. Krackhardt has shown

that the Simmelian ties are more powerful and persistent than the ties that are not simmelian (Krackhardt 1998), but, they come at a price. In a subsequent work he has shown that having Simmelian ties with multiple cliques constrains a person's behavior as he has to conform to norms of several groups (Krackhardt 1999). In a recent work Krackhardt et al. has shown that Simmelian ties positively affect the responses in online discussion forums (Krackhardt 2008).

In this work we examine the effect of such conditions on the reply ties over the blog networks. In addition we also explore the effect of membership in multiple cliques on the strength of the tie. The literature suggests two contrasting effects of clique membership. While being part of a clique strengthens the tie between a pair of actors, being part of multiple cliques *stresses the individual actor*. We hypothesize that although being embedded in a clique strengthens the tie between a pair of actors, when they are embedded in increasing number of cliques the tie between them would be stressed by multiple group norms and thus be *weakened*. However, the nature of the stress in the case of blogs would be different from the nature of stress exerted by cliques in an organization in (Krackhardt 1999). While in the later the actor is stressed because his behavior has to conform to the norms of different groups, in the former the stress will originate from the limited amount of time each has to keep abreast with the topics of interest represented by each cliques in the blogosphere they are part of.

4 Analysis and discussion

4.1 The relative effect of brick-and-mortar factors on online tie formation

In the first two analyses the dependent variables are the citations and replies in the second period. So, for any two actors x and y, they are:

1. Number of replies from x to y in t_2 (rep_{t_2})
2. Number of citations made by x to y in t_2 (cit_{t_2})

The objective is to find out whether the brick-and-mortar ties have a strong influence on the online ties and thus stifle their formation on the blog network. The literature in homophily suggests this might be the case (McPherson, Smith-Lovin et al. 2001). However, the effect of homophily has primarily been studied in the absence of online communication network. The degree to which the ties in the brick-and-mortar world affect formation of new ties in the online world is a largely unstudied problem. In order to measure their effect vis-à-vis the effect of online ties we use two groups of explanatory variables: the brick-and-mortar ties, and the online ties from the first period.

A. Brick-and-mortar ties

We consider the following ties because as described in Section 3.1 they represent strong ties over the brick-and-mortar network. Thus, if there is any effect of homophily from the traditional network on the formation of blog-ties then the following brick-and-mortar tie variables should reveal them.

3. Reporting-to: reporting-to(x,y) = 1 iff x reports to y. (b)
4. Reporting-sibling: reporting-sibling(x,y) = 1 iff x and y have the same boss. (ba)
5. Common-location: common-location(x,y) = 1 iff x and y go to same office. (loc)
6. Common-role: common-role(x,y) = 1 iff x and y have the same designation. (da)

Over the period of data collection the brick-and-mortar ties were time invariant.

B. Online ties in t_1

7. Number of replies from x to y in t_1. (rep_{t_1})

8. Number of citations from x to y in t_1. (cit_{t_1})

9. Number of blogs at which x and y have replied together in t_1. (cc_{t_1})

The explanatory variables are summarized in Table 2.

	Directed	Undirected

Brick-and-mortar ties	`Reporting-to`	`Reporting-sibling`
		`Common-location`
		`Common-role`

Online ties	*Reply* *Citation*	*Reply-together*

Table 2 Explanatory variables: Variables in `fixed font` are dichotomous and those in *italics* are count variables.

4.1.1 Pair-wise correlation among the variables

We start by observing the correlation among all ties (Table 3).

	Correlation among network variables	Boss	Reporting sibling	Common location	Common designation	Citation in t_2	Citation in t_1	Co-comment in t_1	Reply relation in t_2	Reply relation in t_1
Brick-and-mortar ties	Boss	1	0.001	0.011	-0.002	*0.003*	*0.006*	*0.001*	*0.004*	*0.004*
	Reporting sibling	0.001	1	0.142	0.071	*0*	*0.001*	*0.001*	*0.02*	*0.006*
	Common location	0.011	0.142	1	0.115	*0*	*0*	*0*	*0.011*	*0.004*
	Common designation	-0.002	0.071	0.115	1	*0*	*0*	*0.004*	*0.006*	*0.004*
Online ties	Citation in t_2	0.003	0	0	0	*1*	***0.201***	***0.117***	***0.16***	***0.099***
	Citation in t_1	0.006	0.001	0	0	*0.201*	*1*	***0.215***	***0.081***	***0.265***
	Co-comment in t_1	0.001	0.001	0	0.004	*0.117*	*0.215*	*1*	***0.226***	***0.405***
	Reply relation in t_2	0.004	0.02	0.011	0.006	*0.16*	*0.081*	*0.226*	*1*	***0.362***
	Reply relation in t_1	0.004	0.006	0.004	0.004	*0.099*	*0.265*	*0.405*	*0.362*	*1*

Table 3: Correlation among the dyadic variables in two time periods.

Observation 1: Weak correlation between brick-and-mortar ties and online ties.

This suggests that the blog network has considerable independence from the traditional brick-and-mortar network. This does not prove that homophily irrelevant in online tie formation.

Rather, homophily with respect to the brick-and-mortar attributes such as designation, reporting, office location does not limit the online ties.

Observation 2: Positive correlation between online ties.

Online ties have stronger correlation among themselves than with brick-and-mortar ties. Thus the set of blog-ties are cohesive and independent from other ties in the organization. Moreover the online ties formed in time period t_2 have the highest correlation with the similar ties in the previous time period t_1, suggesting that these ties persist over time.

4.1.2 Two period dyadic regression predicting future tie formation

Multiple regressions are the popular choice for studying the effect of several variables on a dependent variable in datasets consisting of records that are independent and identically distributed. However, when each record consists of measurements over a dyad in a social network, the records have structural autocorrelation, hence, cannot be considered independent. In such a case although OLS estimates of the coefficients are unbiased, the estimates of the standard errors will be biased. Therefore, we cannot perform the standard hypothesis testing procedures to infer whether the coefficients are significant. Recently, Dekker et al has suggested a Quadratic Assignment Procedure based approach to estimate the significance of the results (Dekker, Krackhardt et al. 2007). In this approach, regression coefficients are estimated multiple times with explanatory variables permuted. The significances of the original estimates are determined from their position in the resulting empirical distribution of estimates. We use this routine as implemented in UCINET for all our analysis. The first two regression equations are:

$$cit_{t_2} \sim cit_{t_1} + rep_{t_1} + cc_{t_1} + ba + b + loc + da$$
$$rep_{t_2} \sim cit_{t_1} + rep_{t_1} + cc_{t_1} + ba + b + loc + da$$

4.1.3 Results

Variables	cit_{t_2}		rep_{t_2}	
	Raw coefficients	Standardized coefficients	Raw coefficients	Standardized coefficients
ba	-0.000041*	-0.0006*	0.0506*	0.0169*
b	0.001449*	0.0021*	0.0683*	0.0024*
da	-0.000008*	-0.0008*	0.0011*	0.0026*
loc	0.000009	0.0003*	0.0082*	0.0074*
cit_{t_1}	0.0809*	0.1796*	-0.5208*	-0.0280*
cc_t	0.0030*	0.0683*	0.1777*	0.0990*
rep_{t_1}	0.0003*	0.0240*	0.1833*	0.3296*

Table 4 brick-and-mortar ties vs online ties

The estimated regression coefficients are given in Table 4. The starred coefficients are significant at a p-value of 0.05 or less. We can see that the brick-and-mortar ties have much smaller effect than the online ties. The coefficients of brick-and-mortar ties are zero or near zero. Also, consistent with the earlier correlation observations the online ties are most affected by the past online ties of the same kind. However, two additional interesting results emerge.

4.2 Interaction between citation, reply and co-commenting

Observation 3: After the prior reply habit, co-participation in blog discussions has the strongest effect on future reply relations

This indicates that a blog not only act as a communication medium between the blogger and the readers who comment, it also acts as a medium where commenting readers interact leading to future reply relations among them. This transformation of co-commenting tie to reply tie is illustrated in Figure 2.

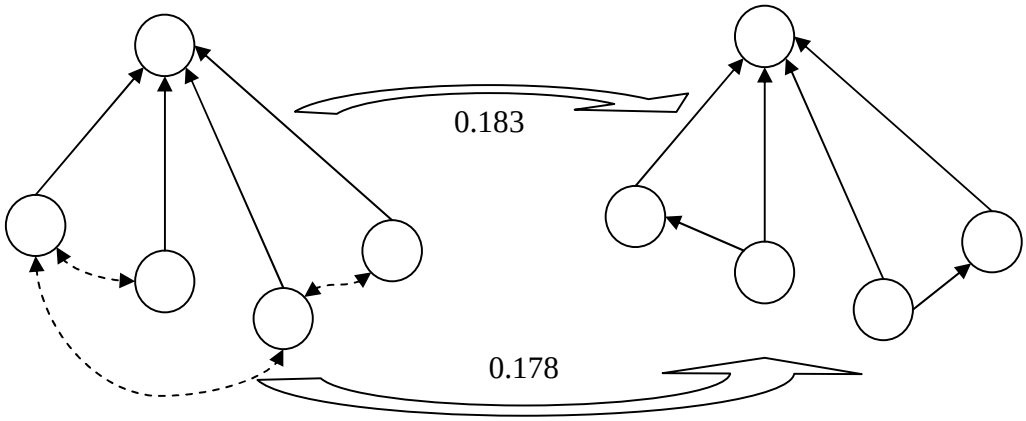

Figure 2 conversion of co-comment relation to reply relation

Observation 4: The citation tie in t_1 has a negative effect on the reply ties formed in time period t_2 while controlling for other offline and online factors

The observed negative effect of citations in t_1 on replies in t_2 is part of an interesting dynamics between citation and reply. Note that in Table 3 we saw a positive pair-wise correlation between the two. This suggests a suppressor effect that is hiding the true correlation between cit_{t1} and rep_{t2}. By sequentially adding the control variables we identify that rep_{t_1} is the variable that is suppressing the true effect of cit_{t_1} on rep_{t_2}. rep_{t_1} positively influences both cit_{t_1} and rep_{t_2} (Table 5).

Explanatory variable	Co-efficient of cit_{t1}
cit_{t_1}	1.5
$cit_{t_1} + rep_{t_1}$	-0.31
$cit_{t_1} + cc_{t_1}$	0.62
$cit_{t_1} + cc_{t_1} + rep_{t_1}$	-0.52

Table 5 rep_{t_1} has a suppressor effect

Therefore, not accounting for its effect would lead us to conclude that cit_{t_1} has a positive effect on the level of rep_{t_2}—quite opposite of the true effect. These two effects are illustrated by a series of linear fits between cit_{t_1} and rep_{t_2} at each level of rep_{t_1} in Figure 3.

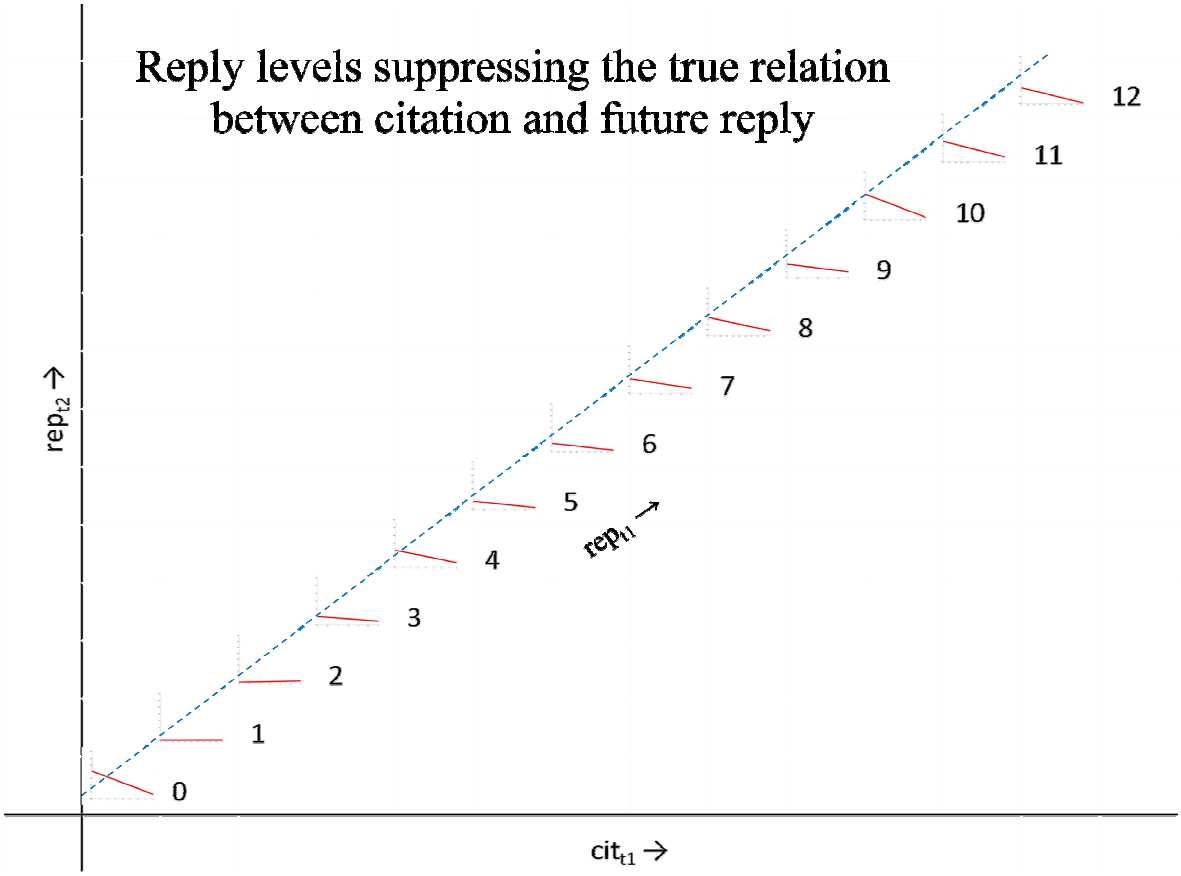

Figure 3: The dotted blue line shows the unconditional correlation between citation and future reply. The little solid red lines show the partial correlation at each level of the current reply.

One theory to explain this is: a reply relation is a conversational relation with the blogger. However, a citation brings notice to a post. Since, citations can usually be thought of as recognition of authority they are facilitated when there is a gap in status of pair of actors. However, this gap in status makes it unlikely for conversational reply relations to develop. Another related explanation is that by citing a post the blogger merely shifting the discussion to his/her own blog. Thus the citations are substituting for the replies. The reason behind this could

be several including a gap in the actors' position in the organization network that inhibits the development of a conversational tie. However, these remain as open research questions.

4.3 Effect of reciprocal and Simmelian conditions on online link formation

As the effect of brick-and-mortar ties turn out to be small compared to the effect of online ties on the future online ties, we focus only on the online ties in this set of analysis. Primarily we explore the effect of reciprocated reply ties and the Simmelian reply ties on the future reply ties. The regression equation we estimate is:

$$rep_{t_2} \sim cit_{t_1} + cc_{t_1} + rep_{t_1} + rr_{t_1} + rs_{t_1}$$

where, rr_{t_1} = indicator variable for the presence of a reciprocated reply tie between x and y in t_1 and rs_{t_1} = indicator variable for the presence of a Simmelian reply tie between x and y in t_1.

We exclude the citation ties from this analysis because citation is much less apparent to the person cited than the reply is to the person replied to. The blogger y is not necessarily aware of the citation made by blogger x, unlike a reply where blogger y sees the reply at this blog. Since, much of the strength of a reciprocated tie is derived from both the party being aware of the reciprocity citation tie is not a good candidate for examining these two conditions.

Predictors	Raw Coefficients	Standardized Coefficients
cit_{t_1}	-0.5764*	-0.0310*
cc_{t_1}	0.1515*	0.0845*
rep_{t_1}	0.1733*	0.3116*
rr_{t_1}	0.1368*	0.0055*
rs_{t_1}	1.3543*	0.0442*

Table 6 Reciprocated ties and Simmelian ties

The estimated coefficients are given in Table 6. The presence of a reciprocated reply tie has a positive effect over and above the effect of prior reply ties. But, the effect of a Simmelian tie on future reply ties is stronger.

As mentioned earlier x and y have Simmelian tie between them if they are part of a clique. It has been pointed out that having more members in the clique does not affect the tie as much as adding one node to a dyad to make it a triad. But, the number of cliques that a pair is part of affects the tie between them. As we argued before, membership in different cliques with different topics of interest impose demands on the individual's limited time. This demand is per topic demand and not per individual contact. In order to show the difference we control for the effect number of friends on a person's activity level.

Predictors	Raw coefficients	Standardized coefficients
cit_{t_1}	-0. 4506*	-0. 0242*
cc_{t_1}	0. 1703*	0. 0949*
rep_{t_1}	0.1782*	0. 3205*
rr_{t_1}	0. 8925*	0. 0355*
nrp_{t_1}	0.0006*	0.0092*
nc_{t_1}	0. 1794*	0. 0236*
$nc_{t_1}^2$	-0. 0419*	-0. 0575*

Table 7 Simmelian tie through many cliques can have a negative effect on the future reply level. Where, nc_{t_1} is the number of cliques x and y are part of in the network of reciprocated reply ties in time period 1, and nrp_{t_1} is the number of people x has replied to in that period.

We find that although, the presence of a Simmelian tie has a positive effect on future reply ties, x and y belonging to more than one clique has a reduced, even negative, impact on future reply

level (Table 7). The relation between Simmelian tie strength and future reply level is non-linear with a negative co-efficient for the higher order term. This is consistent with our hypothesis.

5 Summary

Intra-organizational blogs play an important role as auxiliary social networks. We explore the factors affecting two ties that are the building blocks of this network: citation and reply. We show that the impact of the brick-and-mortar factors on these link formation is minimal compared to the online factors. The online communication channels in the blog network are cohesive and exhibit considerable independence from the brick-and-mortar factors. This by no means is a demonstration of irrelevance of Homophily in tie formation—only that over the blog network people seek out others without being constrained by their brick-and-mortar ties.

Then we explore the online ties that influence the citation and reply tie formation. We find that prior like-ties have the strongest effect, but, beyond them there are interesting interactions among the online ties, such as, the co-commenting ties transform to direct reply ties, citation ties have a negative effect on reply ties when controlled for the prior reply levels, positive effect of reciprocated ties and even stronger positive effect when they are Simmelian. This sheds new light on the nature of the ties in the blog network. However, it also opens new research questions. Some pairs in the blog network might be substituting citation with replies or other way round. Understanding the conditions under which each happens will give us new insights into the nature of citation and reply ties. Being Simmelian with respect to many cliques has a negative effect on the level of reply between a pair. Determination of the mechanism responsible for such a phenomenon will help us better understands the reply networks in the blogs.

Bibliography

Dekker, D., D. Krackhardt, et al. (2007). "Sensitivity of MRQAP Tests to Collinearity and Autocorrelation Conditions." Psychometrika **72**(4): 563-581.

Krackhardt, D. (1987). "QAP partialling as a test of spuriousness." Social Networks **9**(2): 171-186.

Krackhardt, D. (1998). "Simmelian ties: Super strong and sticky." Power and Influence in Organizations: 21–38.

Krackhardt, D. (1999). "The ties that torture: Simmelian tie analysis in organizations." Research in the Sociology of Organizations **16**: 183-210.

Krackhardt, D., Krishnan, R., Kumar, V. (2008). User generated contributions to enterprise wide forums. Fourth Symposium on Statistical Challanges in Electronic Commerce Research.

Liben-Nowell, D. and J. Kleinberg (2007). "The Link Prediction Problem for Social Networks." Journal of the American Society for Information Science and Technology **58**(7): 1019-1031.

Lise, G. and P. D. Christopher (2005). "Link mining: a survey." SIGKDD Explor. Newsl. **7**(2): 3-12.

McGlohon, M., J. Leskovec, et al. (2007). Finding Patterns in Blog Shapes and Blog Evolution. International Conference on Weblogs and Social Media, Boulder, CO.

McPherson, M., L. Smith-Lovin, et al. (2001). "B IRDS OF AF EATHER: Homophily in Social Networks." Annual Reviews in Sociology **27**(1): 415-444.

Ojala, M. (2005). "Blogging: For knowledge sharing, management and dissemination." Business Information Review **22**(4): 269.

Simmel, G. (1950). The sociology of Georg Simmel, Free Press.

Wasserman, S. and K. Faust (1994). Social Network Analysis: Methods and Applications, Cambridge University Press.

6 Figures

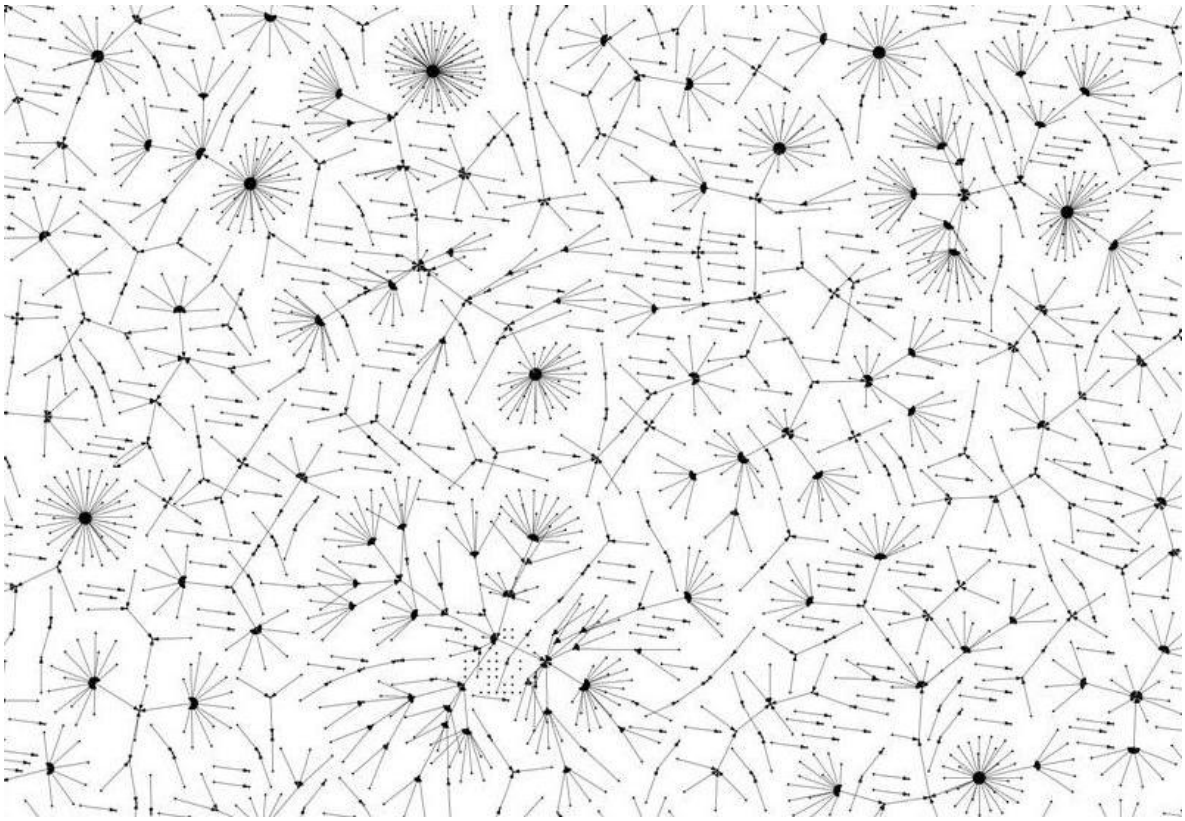

Figure 4 Crop of the reporting network present in the dataset. The directed arrows show the reporting direction.

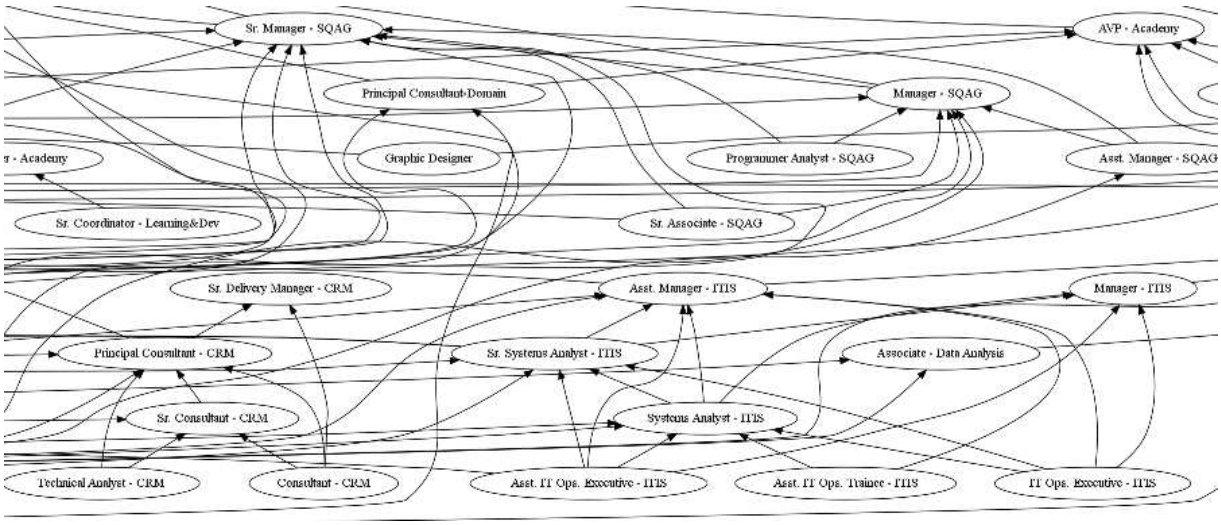

Figure 5 Crop of the role network present in the dataset. The arrows represent the reporting direction of the roles.