Communities on the Web: Mechanisms Underlying the Emergence of Online Discussion Networks

Andreas Kaltenbrunner[1]
Sandra Gonzalez-Bailon[2]
Rafael E. Banchs[1]
[1]Barcelona Media – Innovation Centre, [2] Oxford Internet Institute

ABSTRACT

This paper focuses on the mechanisms that underlie the emergence of different types of discussion networks. Previous research showed that discussion threads are significantly longer and wider when certain topics, like politics, are being discussed. We consider the self-selection of users as one possible mechanism for the emergence of these structural differences. If users decide to take part in some discussions but not others, thus specialising in certain topics, their differences might lead to the emergence of different networks. We do not find conclusive evidence for the self-selection argument, but we do find time differences across topics: users systematically invest more time in certain discussions (i.e. those producing longer threads). Our analyses strengthen the original finding that it is the topic of the conversation, not the characteristics of those participating, what drives the emergence of certain types of networks: some topics invite users to keep conversations alive for a longer period of time. This, we conclude, suggests that users engage in different patterns of interaction depending on the context, much in the same way as different roles are played in different social settings. Future lines of research are considered.

1. INTRODUCTION

The analysis of online interactions can help us shed light into the mechanisms that underlie collective dynamics (Watts 2007). Recent research has focused on email interactions to investigate social phenomena like the diffusion of information, the self-organization of open-source communities, or the evolution of social networks within an organisational structure (Liben-Nowell and Kleinberg 2008, Valverde and Sole 2007, Kossinets and Watts 2006). Web technologies like blogs have also been analysed to uncover the ways in which, through the selection of links, bloggers promote a plural access to politically relevant information (Adamic and Glance 2005; Drezner and Farrell 2008; Hindman 2009). Discussion forums provide yet another setting where online communities have been studied (Fisher et al. 2006; Welser et al. 2007; Adamic et al. 2008). Here interactions arise in the form of threaded conversations: participants exchange messages and, in the process, they create connections much in the same way as emails create links between a sender and their recipients.

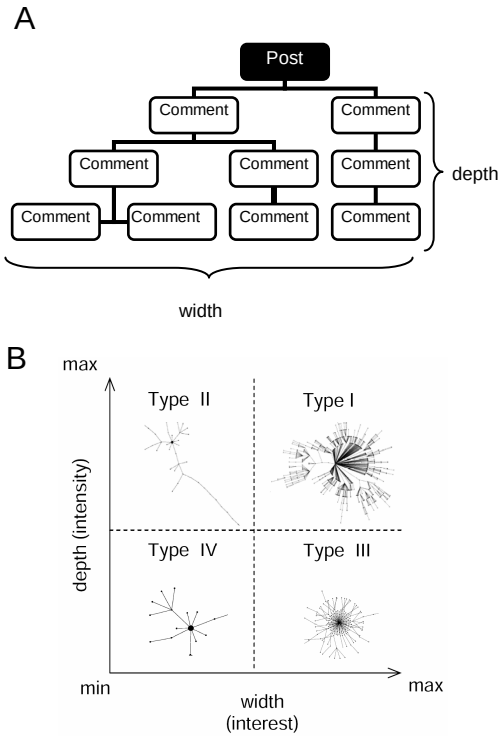
The networks formed by this exchange of messages provide relevant data about how information flows in a particular community, and about the dynamics that drive its emergence. The analyses of online communities can open valuable insights for social scientists, who have long considered networks as relevant contexts to explain individual behaviour (Granovetter 1978; Leighley 1990; Huckfeldt and Sprague 1995; Friedkin and Johnson 1999; Zuckerman 2005; McClurg 2006; Klofstad 2007). The mechanism assumed by this literature is that networks contribute to unleash processes of social influence by shaping exposure to information. What the research on online networks suggests to this literature is that the process of social influence very much depends on the structure exhibited by the networks, which in turn responds to different mechanisms working at the individual level. Previous research shows that users follow different patterns of interaction in different contexts (Welser et al. 2007; Adamic et al. 2008). Differences in the structure of discussion networks might result from this differential behaviour. This paper tries to identify some of these mechanisms, of particular relevance if we are to understand how discussion networks affect exposure to information.

2. STRUCTURAL DIFFERENCES IN ONLINE DISCUSSION NETWORKS

If networks affect individual behaviour by shaping exposure to information, surely the structure exhibited by those networks has a crucial explanatory role. However, the literature linking discussion networks with political behaviour (for instance, Friedkin and Johnson 1999; McClurg 2006; Klofstad 2007) does not usually go into structural properties other than the size or density of ego's connections. In previous research we looked at the structure of more than ten thousand discussion networks and found that their structural properties (and therefore the type of dynamics generated by the discussions) vary significantly depending on the topic being discussed (Gonzalez-Bailon et al. 2008). We analysed the networks formed in the online forum Slashdot from August 2005 to

September 2006, which includes about ten thousand threads, two million comments, and ninety-four thousand unique users. More details about the dataset can be found in (Kaltenbrunner et al, 2008). We focused on two structural features, width and depth, illustrated in panel A of Figure 1. Using these two features, we hypothesized the existence of four types of networks, depicted in panel B: discussions generating higher levels of interest and intensity would materialise in wider and deeper networks, whereas discussions unable to incite many contributions would produce networks both narrower and more superficial. We found a significant association between the topic being discussed and the type of networks generated by the discussion: discussions about politics, for instance, generated networks that were wider and deeper than the average discussion (networks of type I); and discussions about games, generated networks that were narrower and more superficial (type IV).

Figure 1. The Structure of Discussion Networks



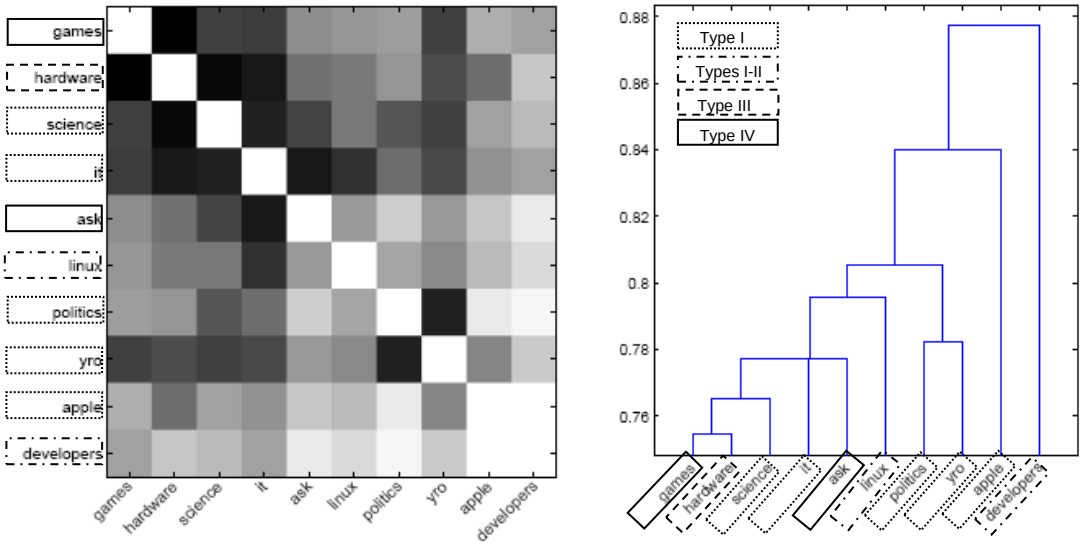
The main conclusion of the study was that not all discussion networks promote the same type of exchange of information. This is relevant for the literature linking networks with political behaviour because it qualifies the effects that discussion networks can have for those participating in the discussions: wider and deeper structures will allow participants to engage in a richer exchange. This poses the question of what drives the emergence of the different structures. In this paper we consider the self-selection of users as one possible mechanism. If users discriminate the topics in which they are willing to participate on the basis of their (unmeasured) attributes or expertise, as suggested by previous research (Adamic et al. 2008), then user specialisation might be the reason why different topics generate different networks. On the other hand, previous research also suggests that users might participate indiscriminately in different discussions, regardless of the topic, but they could still leave different 'structural signatures' in their network interactions when discussing in different settings (Welser et al. 2007). These signatures would be the equivalent of roles, that is, people behaving according to different patterns or expectations depending on the context in which they carry out their interactions (Fisher et al. 2006). These signatures would compose, when aggregated, the different network structures that we identified in the Slashdot discussions. The analyses presented in the following two sections aim to discriminate between these two possible explanations.

3. OVERLAPPING OF USERS

Differences between types of networks might derive from the different profiles of the discussants involved: if users specialise in certain topics, the differences in the types of networks could arise from their different characteristics and, in particular, from their unequal propensity to engage in longer discussions. In addition, if certain users decide not to participate in certain discussions, these would be restricted to a smaller group of specialists, which would undermine the pool of participants from where networks acquire their width. To test for this possible self-selection, we created a participation vector for every major category of discussion as listed in Slashdot when the data was collected. Every dimension of the vector corresponds to a certain user and its value corresponds to the number of times this user participated in the discussions of the given category. The vectors were then renormalized using the tf-idf weighting (Salton 1983). Figure 2 plots the cosine distance of this renormalized participation vectors both as a matrix of overlapping users and as a dendrogram. The legend identifies the type of network that discussions generate, on average, in each category.

Darker cells in the matrix indicate a higher proportion of users shared by the two corresponding categories (diagonal cells are excluded). As the figure shows, some discussion topics like yro (short for 'your rights online') and politics are more similar in terms of users' participation than others, for instance politics and developers: 67% of the users post at least twice under politics and 'yro' but only less than half (31%) participate both in political and developers discussions. However, the dendrogram does not reveal any clustering of categories that clearly maps onto the four types of networks: the two categories that are closer in terms of shared users, games and hardware, generate networks of type IV and III, respectively. Politics and 'your rights online' share many users, but other discussions of type I, like science or it, are closer to games or hardware. In general, the figure shows no clear trend in the overlapping of users that matches the different types of structures that their interactions generate. This poses the question of whether users devote time to the discussions differentially, that is, depending on the category in which

Figure 2. Overlapping of Users across Discussion Topics and Cosine Distance



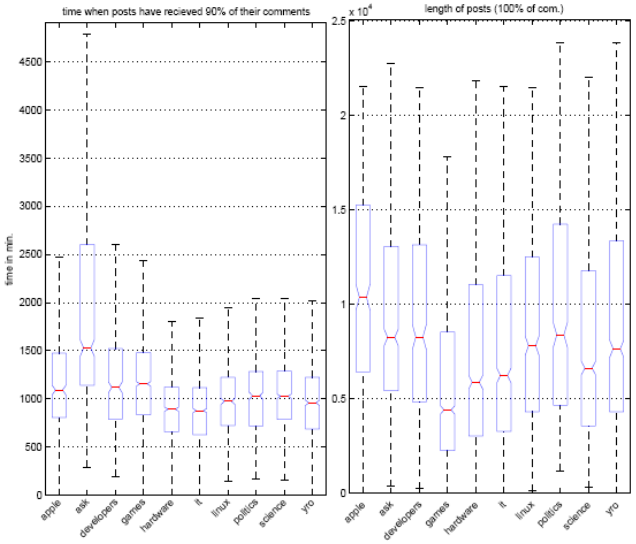
they participate. The following section explores this possibility.

4. TIME PATTERNS

4.1. Global Time Dynamics

The previous section offered evidence against the possibility that structural differences can be explained in terms of user specialisation: categories sharing the highest proportion of discussants exhibit different network structures, which means that in spite of having in common a significant share of users, the emerging networks of interaction follow different patterns. This section analyses the differences in the time patterns generated by the discussions: if threads in certain categories last significantly longer, this might be related to longer paths at the structural level. Figure 3 plots the distribution of thread length, as measured by the number of minutes took by the discussion, across categories:

Figure 3. Quantile Distribution of Thread Length

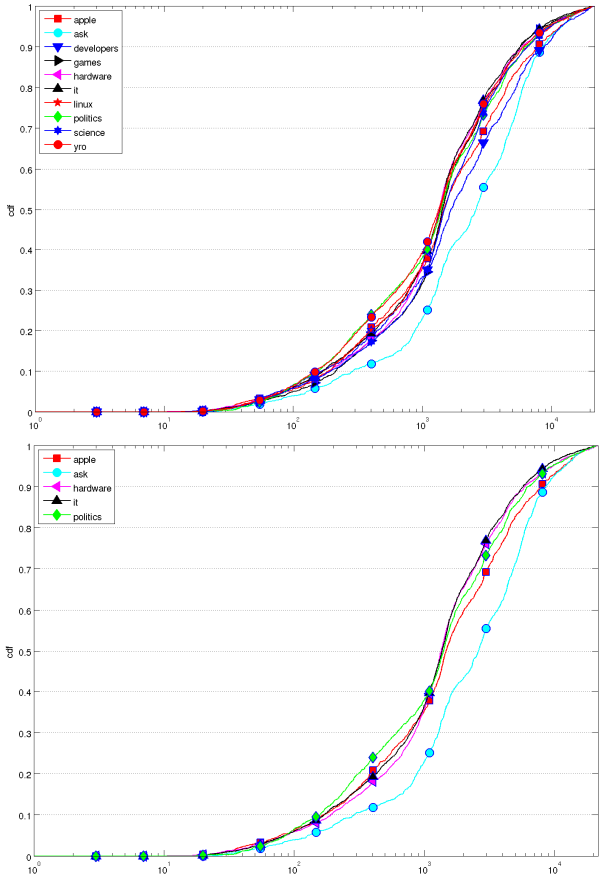


As the contrast between the first and the second graph shows, the final distribution is very much determined by a minority of comments (those in the last 10% quantile) which contribute to lengthen discussions in all categories. However, the graph also shows interesting differences in the variance of length across all topics of discussions: discussions of type I seem to last longer than other types. Given that users do not seem to specialise in certain discussions (which might lead to a concentration of the most engaged users in categories producing networks of type I), these time differences at the global level might result from users following different time strategies depending on the topic being discussed. The following section report analyses of data at the individual level.

4.2. Individual Time Dynamics

In order to identify differences at the individual level – namely whether users systematically follow different time patterns in their participation depending on the topic of the discussion— we selected the subset of users that participate in all 10 major categories. Although it only contains 4.8% of all users, this subset corresponds to the core of Slashdot users, that is, the most active. For every one of these users, we took the posts where they commented at least five times, we calculated the time intervals between their first and last comment (FLI) per post, and we then grouped the FLIs by category. We focused on posts with at least 5 comments to select those with a certain level of engagement without discarding too many, obtaining a total of 25,456 FLIs. The resulting FLI-distributions, displayed in Figure 4, show some interesting differences in the dynamics of discussions. Political discussions, for instance, have more users with shorter FLIs than the other categories (which is a sign of higher engagement) but they also have more users with longer FLIs. Other categories (like discussions about hardware or IT) gain in the intermediate region.

Figure 4. Duration of Participation of Users (FLIs, all and selected categories)



In order to assess the statistical significance of these differences, we used pairwise ANOVA tests comparing the logarithmic transformation of the FLI distributions. Although the means of the entire distributions are very similar for many categories (with high p-values), comparing only the first and the last 50% of the FLIs yields significant differences, as Table 1 shows. This means that users consistently change their reaction times depending on the topic of the discussion, and that it is in the extremes of the distributions where the most significant differences are. The categories "apple", "politics" and "yro" (the three of them generating networks of type I, that is, wider and deeper than the average network) generate discussions with faster dynamics than other types of discussions; discussions in the "developer" category (on the fringes of networks of type I and type II), last for longer, maybe because of the higher degree of expertise required to contribute.

5. SUMMARY AND FUTURE WORK

This paper looked into the mechanisms that might contribute to explain the different network structures that emerge when discussion topics change. We do not find conclusive evidence for the self-selection of users, which suggests that different patterns of interaction arise even when the same users are involved and, vice versa, that the same patterns of interaction can arise even when different users are taking part, as long as discussions share the same topic. Our results show that the topics of the discussions are related with the time that users invest in those discussions. First, longer conversation threads tend to span over longer periods of time; and second, certain topics generate more intense periods of participation among the same subset of users. This might suggest that users adapt their patterns of interaction according to the setting in which their interactions take place, in this case determined by the topic of the discussion. Unlike previous research on the effects that social roles have in discussion dynamics, we link the notion of structural signature to the properties of global networks: when users change the logic of their interactions, that affects their personal networks, but also the structure of entire systems of interaction, as depicted by the four types of discussion networks from which the paper departed.

As future research we intend to improve our understanding of phenomena in online discussion by studying the time evolution of the generated thread structures. We will furthermore analyse the individual interaction patterns of single users in the threads via their local ego-networks. This will allow a more detailed characterisation of those individual and global patterns

Table 1. p-Values of pairwise ANOVA tests for FLI time distributions of users participating in all categories (values for the first 50% of time intervals in grey, for the last 50% in white cells; bold numbers are significant at the 5% level)

	Apple	Ask	Developers	Games	Hardware	IT	Linux	Politics	Science	YRO
Apple		**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**
Ask	**.00**		**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**
Developers	.20	**.00**		**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**
Games	**.00**	**.00**	.21		.40	**.04**	.54	.06	**.00**	.86
Hardware	.20	**.00**	.76	**.03**		.28	.22	**.01**	**.00**	.50
IT	.86	**.00**	.19	**.00**	.17		**.03**	**.00**	**.00**	.07
Linux	.66	**.00**	.34	**.00**	.40	.74		.33	.06	.49
Politics	.05	**.00**	**.00**	**.00**	**.00**	**.01**	**.01**		.40	.06
Science	**.01**	**.00**	.38	.61	.09	**.00**	**.02**	**.00**		**.00**
YRO	**.01**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	.57	**.00**	

within the discussions, and will provide means for a comparative study of behaviour in different discussion topics. Additionally, the study of temporal evolution will allow us to characterize discussions by their trajectory across the width depth plain and explain for example how a discussion evolves from type IV to type I. Ultimately, this line of research will allow us to generate a simple theoretical model of thread evolution, where a few parameters allow us to describe the shape and growth of online discussions.

In addition, we also plan to provide a more detailed account of the mechanisms that explain the emergence of different types of networks by looking into the content of the messages exchanged. We want to test if we can explain the faster dynamics of certain discussions in terms of the emotional engagement that certain topics instil in the participants. We plan to conduct quantitative content analysis of the messages exchanged to see if certain words of high emotional content appear more frequently in deeper and wider networks than in other types of discussions. We also want to test if reciprocity, as measured by number of messages exchanged between the same two participants, is systematically higher in some types of networks. This will allow us to link network structure with the existence of absence of social norms.

ACKNOWLEDGEMENTS

This work has been partially funded by Càtedra Telefònica de Producció Multimèdia.

REFERENCES

Adamic, Lada, and Natalie S. Glance. 2005. "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog." in *2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, WWW 2005*. Japan.

Adamic, Lada, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. 2008. "Knowledge Sharing and Yahoo Answers: Everyone Knows Something." in *WWW 2008*. Beijing, China.

Drezner, Daniel W., and Henry Farrell. 2008. "Blogs, politics and power: a special issue of Public Choice." *Public Choice* 134:1-13.

Fisher, Danyel, Marc Smith, and Howard T. Welser. 2006. "You are Who You Talk To: Detecting Roles in Usenet Newsgroups." *HICSS'06*.

Friedkin, N, and E. Johnson. 1999. Social Influence Networks and Opinion Change. *Advances in Group Processes*, 16: 1-29.

Gonzalez-Bailon, S., Kaltenbrunner, A. and Banchs, R. 2009. “The Structure of Political Discussion Networks: A Model for the Analysis of E-Deliberation.” *Under Review*.

Granovetter, M. 1978. Threshold Models of Collective Behaviour. *American Journal of Sociology*. 83(6): 1420:43.

Hindman, Matthew S. 2009. *The Myth of Digital Democracy*. Princeton, NJ: Princeton University Press.

Huckfeldt, R. and J. Sprague. 1995. *Citizens, Politics, and Social Communication: Information and Influence in an Election Campaign*. Cambridge: CUP.

Kaltenbrunner A., Gómez V., Moghnieh A., Meza R., Blat J., López V. 2008. Homogeneous temporal activity patterns in a large online communication space, IADIS International Journal on WWW/INTERNET, 6(1): 61-76.

Klofstad, C. 2007. Talk Leads to Recruitment: How Discussions about Politics and Current Events Increase Civic Participation. *Political Research Quarterly*. 60(2): 180-91.

Kossinets, Gueorgi, and Duncan J. Watts. 2006. "Empirical Analysis of an Evolving Social Network." *Science* 311:88-90.

Leighley, J.E. 1990. Social Interaction and Contextual Influences on Political Participation. *American Politics Research*. 18(4): 459-75.

Liben-Nowell, David, and Jon M. Kleinberg. 2008. "Tracing Information Flow on a Global Scale using Internet Chain-Letter Data." *PNAS* 105:4633-38.

McClurg, S. 2006. The Electoral Relevance of Political Talk: Examining Disagreeement and Expertise Effects in Social Networks on Political Participation. *American Journal of Political Science*. 50(3): 737-54.

Salton, Gerard, 1983. *Introduction to Modern Information Retrieval*. Mcgraw Hill Computer Science Series.

Valverde, Sergi, and Ricard V. Solé. 2007. "Self-organization versus hierarchy in open-source social networks." *Physical Review E*:046118-1-8.

Watts, Duncan J. 2007. "A twenty-first century science." *Nature* 445:489.

Welser, Howard T., Eric Gleave, Danyel Fisher, and Marc Smith. 2007. "Visualizing the Signatures of Social Roles in Online Discussion Groups." *Journal of Social Structure* 8(2).

Zuckerman, Alan S. (Ed.). 2005. *The Social Logic of Politics: Personal Networks as Contexts for Political Behavior*. Philadelphia, PA: Temple University Press.