THE STRUCTURE OF GEORGIAN BLOGOSPHERE
AND IMPLICATIONS FOR INFORMATION DIFFUSION*

Zakaria Babutsidze,1,2,† Tengiz Lomitashvili2 & Koba Turmanidze2,3

1 OFCE Sciences Po, 250 Rue Elbert Einstein, 06560 Valbonne, France
2 Center for Social Sciences, 3 Chavchavadze Avenue, 0179 Tbilisi, Georgia
3 Caucasus Research Resource Center, 16 Zandukeli Street, 0108 Tbilisi, Georgia

Preliminary draft
Do not cite without permission from the authors!

August 5, 2011

Abstract

We present the first study of the Georgian blogosphere. The complete
dataset on Georgian bloggers is collected and analyzed using Social Network
Analysis tools. We identify the core-periphery structure present in the system.
Results point to the fact that, after disregarding least important peripheral
blogs that do not actively participate in the system, Georgian blogosphere is
very compact. The core of the system cannot be partitioned into communities
by any observable characteristics. Therefore, we perform a community detec-
tion exercise. We identify two large communities and analyze their behavior,
internal architecture and inter-community interactions.

1 Introduction

The strive to express own and to influence other people’s opinion is a strong mov-
ing force of social individuals. Democracy that requires social consensus actively
encourages public debate on different platforms.

The creation of internet and its high penetration levels into different types of
societies presents us with the new platform for debate – a blogosphere. The speed

* Authors acknowledge financial support of Academic Swiss Caucasus Network for carrying out
this work. Comments from Giga Zedania and Nicolas Hayoz are also gratefully acknowledged.
† Corresponding author. E-mail: zakaria.babutsidze@sciences-po.fr.
user-friendly blog-hosting websites has contributed towards this a great deal. Today blogging is a popular activity among socially active individuals all across the globe. Besides its popularity, blogging has another big advantage over other discussions domains for researchers. Because of the fact that discussions in blogosphere happen in virtual space (internet) they leave the electronic trace, large amounts of information that can be collected at a later stage and analyzed.

The part of blogging practices that is most interesting for social scientists is political blogging. Groups of (or isolated) individuals can run blogs with the aim to support certain cause/ideology and influence decision-makers or their electorate. The ground for the analysis of political blogosphere has been laid by Adamic and Glance (2005) who investigate blogging behavior of Democrat and Republican blogs before 2004 presidential elections in the United States of America. Adamic and Glance (2005) were interested in differences in linkage patterns of blogs from the two camps. Another fundamental contribution has been due to Blumenthal (2005), who has been concerned by the content generated by the blogosphere, rather than the linkages between the blogs.

Since then there have been many contributions towards the analysis of the political blogosphere of different countries. And the importance of blogging for the different aspects of the social and political life has been continuously demonstrated (Drezner and Farrell, 2008). It has been argued that blogging practices reflect the political and social processes going inside the country. Besides this they have been identified as an important force that can steer these processes. Blogs and a blogosphere (defined as the collection of blogs) can generate information cascades (Leskovec et al., 2007) that can spread into different parts of the society with breathtaking speed (Gruhl et al., 2004). The power of bloggers in influencing public opinion has been documented by measuring sentiment diffusion (Kumar et al., 2003). The background for this research has been provided by the research into the spread of infectious diseases that are usually transmitted through social networks (Bailey, 1975).

But blogosphere is not homogenous. Blogosphere just as any social network consists of communities. These are the collections of blogs that have unified goal. The goal of a community might well be going against the goal of another community. Analytical framework for communities in blogging has been provided by Schmidt (2007). Sometimes, like in the case of Adamic and Glance (2005), communities can be identified by their observable characteristics. Hargittai et al. (2008) investigate the blog linking behavior of conservative and liberal bloggers and find a
strong support for clustering within the communities. They also test whether this high density of interlinkages between the blogs in the same community is becoming stronger under the influence of technological advances in web-search and sentiment mining. However, they conclude that the linking patterns and therefore the partition of blogosphere’s into communities, are stable. This means that the communities do not get further separated from each-other with the passage of time.

There are two major characteristics of the research into blogging up to date. First is that it analyzes only political blogosphere. Political blogosphere is a small part of the whole web that has a deliberate aim of influencing opinion. There are two reasons behind this decisions. First is that these are the most influential blogs. The second reason is that usually whole of the blogosphere of a country is much larger and the data collection on all of them represents much harder problem to solve. The second characteristic of the research into blogging practices is that it concentrates on large western economies (Ali-hasan and Adamic, 2007 that analyze Kuwait and United Arab Emirates is an exception to this).

In face of past research the analysis of small developing countries presents valuable opportunities. Firstly, small and developing countries that do not have yet the overwhelming number of active bloggers. Therefore, the detailed analysis of the entire blogosphere (not only its political portion) is feasible. Besides, by taking this opportunity now the researcher very likely to observe the blogosphere at its infancy. This gives the opportunity for taking a first step in collecting a rich dynamic data about the blogosphere which is potentially going to become a very important player in shaping public opinion in a given country. Secondly, the negligence of the analysis of blogging data in small countries leaves a very important question open: do blogospheres in these small countries behave differently than the ones in large countries?

In this study we dissect and take the first glimpse into Georgian blogosphere. Georgia is a country with high speed of development and social transformation that is taking place as this piece is being written. This transformations are coinciding with the popularization of the blogging behavior. From this prospective, bloggers have a chance of playing a unique, more prominent role in these progressive but painful processes than in other countries. This makes the study of Georgian blogosphere attractive.

Besides, the country choice presents the unique opportunity to collect the complete dataset on Georgian bloggers. Georgia has a unique alphabet that gives a sharp tool to limit the web search. Besides, Georgian blogosphere is overwhelmingly
monolingual. Therefore, the boundaries of the data on Georgian-language blogs co-incide with the boundaries of Georgian blogosphere. This statement is rarely true for other countries which have unique alphabet (for example non-negligable parts of the blogosphere in neighboring Armenia operates in Russian, English or even French languages). This is important as networks, as is the blogosphere, do not have representative samples. Therefore, investigating any part of them, no matter how carefully selected, is bound to bias the conclusions.

As asserted earlier blogging practices are found to be important for opinion formation. Even though number of blogs is relatively small to the population of the countries and furthermore, even the number of people reading blogs is not that large either, people creating and participating in blogosphere are the ones that have power diffuse information and sentiments further, well beyond the borders of the blogosphere. Therefore blogs and activity going around them can be used as relatively accurate barometers of political life in a given country.

However, the results of the studies similar to the one presented in this paper have to be taken with a grain of salt. We do not mean to suggest that links in our data will necessarily pass all sorts of information or that blogs that are linked to each-other will necessarily influence each-others actions. These are the explorations into the potentials of existing social networks. Although Ali-hasan and Adamic (2007) find that significant amount of relations are formed through blogging interactions, recent study bu Aral et al. (2009) documents that the structure of the online social networks is largely influenced by the homophily and that similar action of linked individuals can be a result of this factor rather than one of them influencing the other.

With these caveats in mind we further go to present the analysis of Georgian blogosphere.

2 Data and methodology

The paper utilizes a large dataset on Georgian-language blogs. The data has been collected on December 11, 2010. First we have built a database of Georgian-language blogs. We started from the already existent database of around 600 blogs. Computer algorithm harvested all the URL citations listed on the front pages of these blogs. These include citations existent in last few blog-posts of each blog plus blogs that are listed in a blogroll of each blog. Following Adamic and Glance (2005) we do not discriminate between these two types of URL’s (ones cited in posts and ones included in a blogroll). Further all harvested URL’s were analyzed and determined whether
they were Georgian-language blogs that were not already listed in our database. Blogs that were new were added to the database and algorithm proceeded in collecting similar URL data as it did to the original 600 blogs. And process went on until no new Georgian-language blog was detected. Using this scheme we were able to collect 3273 Georgian-language blogs which gives a complete picture of Georgian blogosphere.¹

After building this database, on December 11, 2010 a new computer robot entered each of the blogs listed in the database and extracted following data from each of the URL’s. We collected all the citations listed on the front page of each blog and in their blogrolls. For the blogs that are listed on Alexa.com we have collected the ranks attributed to them.² Whenever possible we also collected the data on last 5 posts of the blogger. These included the time-stamp of each post, as well as number of comments it generated and number of unique commentators participating in each discussion. We have also noted down the names/usernames of the commentators. Afterwards these names have been matched to our blog database and commentators that are running their own blogs (that clearly are in our blog database) have been identified.

Unfortunately due to the specifics of the blogging platforms the data on comments is very partial. In particular we were not able to extract comments data from the blogs that are hosted by wordpress.com or are using wordpress engine to run the blog at a private host webpage. In contrast to wordpress.com, data has been easily collected from another large blogging service provider blogspot.com. All in all, we were able to collect the data on posts only for 32% of the blogs. Due to the large chunk of data missing from the comments, we were not able to utilize comments data at its full potential.

From the collected data the network of Georgian blogosphere has been constructed. In this network nodes are bloggers and links are citations. This is a directed network where arrow of the link is pointing towards the blog that is cited. Therefore, we can state that information streams in the opposite direction to the links. Of course if blog A cites blog B and blog B cites blog A there will be links both pointing from A to B and from B to A.

The visualization of the complete data on citation network is presented in figure 1

¹Unlike many neighboring countries Georgian blogosphere is overwhelmingly monolingual. Therefore, our algorithm guarantees that we are not omitting any important players or discussions that might be going on between Georgian bloggers.

²Alexa.com attributes a traffic rank to websites based on the reputation scores that they compute based on the last three months of browsing data collected from the users of Alexa toolbars for Internet Explorer, Mozilla Firefox and Google Chrome.
1. The network collects 3273 blogs/nodes and 7987 citations/links between the pairs. As one can directly see that the network consists of heavily connected core and much less connected periphery. There are many standalone blogs that neither cite, or are cited by any other blog. There are small groups that cite only each other but are not connected to the core of the blogosphere. More details about the structure of the network will be given in the following section.

Descriptive characteristics of the collected data are presented in table 1. As it can be seen from the table, Georgian blogosphere is not as active as blogosphere in developed countries (e.g. in USA as reported by Adamic and Glance, 2005). However, there are two points to make here. Firstly, there are the averages over the whole Georgian blogosphere and their comparison to the studies using only partial data that collects only the most active blogs is somewhat problematic. These studies are bound to overestimate the activity as they are selectively observing the

<table>
<thead>
<tr>
<th>Data</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>3272</td>
</tr>
<tr>
<td>Number of links</td>
<td>7987</td>
</tr>
<tr>
<td>Average intensity of posting</td>
<td>0.009</td>
</tr>
<tr>
<td>Average number of comments generated</td>
<td>3.229</td>
</tr>
<tr>
<td>Average number of unique commentators</td>
<td>2.452</td>
</tr>
</tbody>
</table>

Table 1: Some characteristics of the Georgian blogosphere.
most active blogs. Secondly, Georgian blogosphere is extremely heterogeneous, we explore this point later on. Network behavior is usually characterized by extremely skewed distributions (Watts and Strogatz, 1998; Barabasi and Albert, 1999; Border et al., 2000; Wagner and Fell, 2001) for which average values do not carry much of meaning.

We approach the collected data using social network analysis methodology. Using the stat-of-art tools in these fields we try to identify the blogs that occupy key positions in Georgian blogosphere. In particular we use node statistics such as degree centrality, betweenness centrality, page rank etc. We also analyze the community structure in georgian blogosphere. Using unweighted directed network that we have presented in this section and its weighted undirected transformation we identify two major communities which are further explored in detail.³

However, the plot on the figure 1 is enough to state that in contrast to Adamic and Glance (2005) who through blog linkages identify clear limits between Republican and Democrat blogs in US political blogosphere Georgian blogosphere is not split in two or multiple groups. The part of the blogosphere that seems to be still connected than to the core in the left bottom part of the figure 1 is considerably sparser compared to the center of the network. Therefore it cannot be considered as a different independent unit that might be competing with the core for the flow of the information. This speculation is confirmed by the analysis in the following section.

³We do not collect the content of the posts and comments and therefore do not do into opinion mining exercises (Pang and Lee, 2008).
3 Basic characteristics of Georgian blogosphere

In previous section we have noted that in our network the distribution of the number of links per node (degree of a node) is very skewed. This is not a unique characteristics of our network. Skewed degree distributions have been identified in many blogging and other social networks (Travers and Milgram, 1969; Ebel et al., 2002; Newman et al., 2002; Holme et al., 2004). The left panel of figure 2 presents in- and out-degree distributions for the full network. In degree of a blog is the number of other blogs citing it, while out degree of a blog is the number of blogs it cites. As can be seen there large number of nodes have degree 0 or 1 and at the same time there are considerable number of blogs having extremely high in- and out-degrees.

The degree distribution of many real-world networks have been analyzed in detail and it has been found that many of them can be well approximated by the power laws. In order to see how our network fits into this pattern we plot the degree distributions on log-log scale on the right panel of the figure 2. Power law distribution would be a straight line on a log-log scale. As one can see, both in- and out-degrees of Georgian blogosphere do not represent any kind of exception from the general pattern.

In order to make conclusions about the potential speed of information diffusion in Georgian blogosphere it is better to analyze only a subset of our data. Parts of the network that are not connected to the core of the system are isolated from it and do not play a functional role in spread of information. Therefore, we can disregard them and concentrate our analysis on the core of the system. In order to reduce our network we can use extract two types of subnetworks. One is the largest weakly connected component which unites all the nodes that are able to either receive or to pass the information to the core. The other one is the largest strongly connected component, that is a collection of only the nodes that can receive as well as pass the information to the core. Obviously the strongly connected component is the subset of the weakly connected component.

The largest weakest connected component comprises 2098 nodes, which represents 64.1% of the full citation network. The second largest weakest component has only 26 nodes, therefore it and all the other even smaller ones are disregarded. The largest strongly connected component contains 414 nodes, which corresponds to 12.65% of the full network. The second largest strongly connected component is considerably smaller (only 61 nodes) and is ommited from the analyses. In figure 3 we present plots of in- and out-degrees of the largest weakly and strongly connected components. As it can be easily seen the subsets behave very similar to the original
network.

Different characteristics of the full citation network as well as its largest weakly and strongly connected components are given in table 2. The transition from the full network towards the strongly connected component depicts the difference between the full network and its active core. As one can see from the table the core 12.6% of the nodes account for 35.5% of the links. As a result an average node in the core has 2.8 times more links compared to the network average. As a result, graph density increases 17 times as we go from the full network to its largest strongly connected component. This highlights the difference of the core blogs in terms of linkage behavior with regard to the network average.

In order to assess the length of the path a piece of information has to cover while circulating in Georgian blogging network we calculate two statistics: the average shortest path length and the network diameter. The shortest path length between two nodes is the minimum number of intermediate nodes that information has to go through in order to reach one of the nodes while starting from the other. If

<table>
<thead>
<tr>
<th>statistic</th>
<th>full network</th>
<th>weak</th>
<th>strong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>3272</td>
<td>2098</td>
<td>414</td>
</tr>
<tr>
<td>Number of links</td>
<td>7987</td>
<td>7473</td>
<td>2837</td>
</tr>
<tr>
<td>Average degree</td>
<td>4.882</td>
<td>7.124</td>
<td>13.705</td>
</tr>
<tr>
<td>Graph density</td>
<td>0.001</td>
<td>0.002</td>
<td>0.017</td>
</tr>
<tr>
<td>Average shortest path length</td>
<td>–</td>
<td>5.539</td>
<td>3.74</td>
</tr>
<tr>
<td>Network diameter</td>
<td>–</td>
<td>17</td>
<td>10</td>
</tr>
<tr>
<td>Average clustering coefficient</td>
<td>0.113</td>
<td>0.16</td>
<td>0.204</td>
</tr>
</tbody>
</table>

Table 2: Some characteristics of the full citation network and of its largest weakly and strongly connected components.
the shortest path length is small, information is easy to diffuse in the system. The average shortest path length is calculated by considering all the possible pairings in the network. The network diameter is the shortest path length between the two nodes that are the most distant from each other. This measures the longest route information has to travel in order to reach any node starting from any other node. Due to the fact that the full network contains standalone nodes average shortest path length and network diameter are undefined or stand at infinity. This is to say that the shortest route between two nodes is infinitely long, which effectively means that information cannot reach from one node to the other. However, we can see that the strongly connected component is pretty compact. Its shortest average path length stands at 3.74 and the network diameter is 2.4% of the number of the nodes. Of course, these values are affected by the fact that the size of the strongly connected component is small. However, comparable values have been hound in much larger Twitter network studies by Bakhshandeh et al. (2011). Once we go to the weakly connected component the average shortest path length nears to the universally identified number of 6 degrees of separation (Watts, 2003; Barabasi, 2003; Newman et al., 2006). However, the diameter of the network shrinks compared to its size. Now the diameter amounts only to 0.8% of the network size as measured by the number of nodes. Here note that weakly connected component amounts to the 64.1% of the overall nodes. This all points to the fact that Georgian blogging network, or at least connected part of it, is rather compact and well interlinked, that the potential speed of information diffusion is relatively high.

In order to better understand the linkage behavior of Georgian blogs we calculate the clustering coefficient. This measures what portion of the friends of a node (i.e. the other nodes that it is linked to) are also friends of each-other. The figures in table 2 point to the fact that the core of the system is not only denser but also more clustered pointing to a stronger sense of community in the core compared to the average of the network. We take a closer look into the communities in Georgian blogosphere in the following section.

Every node in a network occupies a unique position. There are three major roles that a node can play and the extent to which a given node performs is evaluated by the appropriate measure. The three important node-level measures/characteristics are in-degree, out-degree and betweenness centrality. A node with a high in-degree is a content generator. This is the node that is cited heavily by the other blogs in the system. A node with the high out-degree has a superior position in spreading information (i.e. the content generated by other nodes). A node with high be-
tweenness centrality has a role of an information broker. This is a node that is very valuable for the circulation of the information, in a sense that its removal from the system increases average shortest path length considerably. Table 4 in the appendix presents 10 top ranked blogs in Georgian blogosphere with respect to each measure. As one can easily identify from the table, there is a very little overlap between the three lists. This means that the three roles described are clearly distributed among the different bloggers.

4 Communities in Georgian blogosphere

Adamic and Glance (2005) have analyzed a political blogosphere of the USA. They have collected their data with predefined characteristics of whether blogs were Republican or Democrat. In this case they had two pre-defined communities that were analyzed and contrasted to each other. In Georgia political blogging cannot be separated from the personal blogging. This is the reason why we are handling the data for the entire Georgian blogosphere. Therefore, we cannot assign any specific pre-defined traits to blogs. Thus, we have to judge on their belonging to certain communities from their behavior.

As asserted earlier the visualization of the data reveals that there is a core-periphery structure in the network. However, it does not suggest that there is a clear pattern of separated communities that would look like large number of nodes structured in different places. But as in all the aspects of social behavior, blogging is bound to have groups of more likely-minded people. Fortunately there are algorithms which can identify communities based on the linkage structures of different nodes. In this section we utilize one of these algorithms, identify communities of a significant size and analyze their internal structure as well as interlinkages.

4.1 Community detection

The measure that gives us the signal that the partition into communities is sensible is called modularity (Newman, 2006). Higher modularity scores correspond to the more pronounced community structure. Usually modularity scores above 0.4 are considered meaningful. Because of the fact that mathematically correct partition that would maximize modularity of a given network is computational extremely expensive we are forced to use an approximation algorithm. We use an algorithm proposed by Blondel et al. (2008). This computationally efficient algorithm detects communities in a network simultaneously maximizing modularity of the partition.
However, there is still a seed of randomness involved. The algorithm can return (at least quantitatively) different results in case of different starting conditions. Starting conditions is basically a node that is selected randomly. In order to overcome this disadvantage we ran the algorithm 100 times and each time record the modularity score of a partition. At the end of an experiment we select the modularity score that is highest and work with the community partition corresponding to it.

Using this algorithm we have identified a community structure in our network that corresponds to the modularity of 0.56. The two largest communities in the citation network are highlighted in the largest weakly connected component that is plotted on the left panel of figure 4. On the right panel only the network consisting by the nodes belonging to these two communities is considered. On this panel the size of a node corresponds to its in-degree (in this network): the larger the node size, the higher the node degree. The two components are quite large and comprise 442 and 332 nodes respectively. The third largest component consists of only 37. Therefore, this and all smaller components have been dropped from the analysis in this chapter.

Each of the algorithms calculating communities have their specific disadvantages (Zhao et al., 2011). We have outlined the disadvantage of the Blondel at al. (2008) method that we have used and an aditional step we took in order to overcome this disadvantage. However, just as a robustness check we take one additional step. Recall that the network that we are working is a citation network. It is a directed unweighted network that puts a link between two blogs if one blog cites the other. This is a representation of the blogosphere that is suited to analyze the direction...
of potential information streams. However, besides this data we also have a commentator data. We can identify bloggers and non-blogger readers that comment on posts of each of the blogger. Besides, if a blogger is citing another blog this means that he is also a reader of this blog. Therefore, we can construct the data about the audience of each of the blogger.

Utilizing this data we can transform our directed unweighted network into undirected weighted network in a following way. We consider that there is a link between two blogs if they share the audience. Hence, there is no direction in links in this network. And weights are put on each of the link proportional to the number of shared users. For example if blog A and blog B have 2 shared readers while blog C and blog D have 4 shared readers, we would say that the link between C and D is twice more important than the link between A and B. This transformation puts emphasis at the readers and, under the assumption a blog would share larger number of readers with the blog in its own community then with the blog outside its community, gives us the opportunity to check the robustness of our partition. The result is a much denser weighted network with the number of links increasing from 7987 to 60487.

We run the same algorithm on this transformed network and select the partition with the highest modularity score. The robustness of our original partition is surprising. Recall that in original citation data identified two largest communities having 774 blogs in total (442+332). The robustness check with the transformed network resulted in two largest communities having 745 nodes. Between these two partitions 587 nodes are shared. And among these 587 nodes 530 is classified as the part of the same community by both mechanisms, which means an overlap of over 90%. This means that the two identified communities are distinct not only in terms of linkage behavior but also with respect to their audiences.

Of course there are small differences in between two partitions, but we choose to work with the original partition and network as it depicts the potential information flow structure more accurately. In what follows we investigate the structure of the two communities and interlinkages between them.

4.2 Structure of the communities

Figure 5 presents internal structure of the two communities. On the left panel we have a largest community (community 1), while on the right panel we present the smaller community (community 2). The intensity of the color as well as the size of
Table 3 presents some statistics for the two communities. As it can be seen from the table two communities are sufficiently different from each other. The larger of them, community 1, is sparser. It has lower average degree and higher network diameter compared to its size. In addition it is also less clustered compared to its counterpart. This concerns the in-community linking structure. However, when it comes to the posting behavior community 1 is performs better. Blogs belonging to the larger community blog more intensively. They also spur longer discussions and involve more people into them.

The leaders of the two communities are listed in table 5 in the appendix.

Table 3: Some characteristics of the two largest communities in Georgian blogosphere.

<table>
<thead>
<tr>
<th>statistic</th>
<th>community 1</th>
<th>community 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>442</td>
<td>332</td>
</tr>
<tr>
<td>Number of links</td>
<td>1816</td>
<td>1512</td>
</tr>
<tr>
<td>Average degree</td>
<td>8.217</td>
<td>9.108</td>
</tr>
<tr>
<td>Graph density</td>
<td>0.009</td>
<td>0.014</td>
</tr>
<tr>
<td>Average shortest path length</td>
<td>3.806</td>
<td>3.153</td>
</tr>
<tr>
<td>Network diameter</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Average clustering coefficient</td>
<td>0.126</td>
<td>0.168</td>
</tr>
<tr>
<td>Average intensity of posting</td>
<td>0.01</td>
<td>0.008</td>
</tr>
<tr>
<td>Average number of comments generated</td>
<td>3.366</td>
<td>2.725</td>
</tr>
<tr>
<td>Average number of unique commentators</td>
<td>2.405</td>
<td>2.166</td>
</tr>
</tbody>
</table>
4.3 Linkages between communities

To analyze interlinkages, we can treat a community as one node and consider its position among the nodes of another community. For instance we can consider community 1 (we call this a guest community) as being one additional node in community 2 (a host community). In this case if a node in the host community is citing at least one of the nodes in the guest community there will be a link between this particular node to the collective node of the guest community.

In this way we can analyze relative importance of each of the communities for the other community. However, the approach is not perfect as it does not allow for weighting links between a node in the guest community. For example, the link between a node in the host community and a guest community node that exists because the node is citing one blog in the guest community and the link that exists because this node is citing 10 nodes in the guest community are treated as equal. In principle it is possible to weight these links, but herein we present the analysis with unweighted links.

In this case we have two scenarios to consider. The first where the community 1 is a guest community and the community 2 is a host community, and the second where the roles are reversed. The community one is a guest and is treated as a single node added to the internal structure of community one, the in-degree of a collective guest community node is 93, while the out degree is 97. In order to get the idea of the importance of the guest node we compare its characteristics to the host community nodes with the highest performance with respect to in- and out-degree. In this instance in-degree of 93 of a guest node has to be contrasted to the in-degree of 47 of the best performing node in host community. The contrast is clear, however keep in mind that we are comparing a collective node that in reality comprises 442 nodes to the performance of a single node. With respect to the out degree the difference is much smaller as out-degree of 97 of the guest node has to be contrasted with the out degree of 85 of a single best performing node in host community.

If we consider the second scenario where community 2 is a guest community (with a single node collecting its 332 members) into the community 1 contrasts between the guest and best performing local nodes are significantly higher. In this case in-degree of a guest node is 102 in contrast to the 39 of the best performing node in community 1 and the out degree is 146 in contrast to the best performing node having out-degree of 49.

As a result we can conclude that compared to the internal linkage behavior of
the community, the smaller community 2 is more important for the community 1 than the other way around. This is true in terms of in-degree as well as out-degree linkages. Which effectively means that community 2 plays a more important role as a content generator for the community 1 than the other way around. And it also plays a more important role as the facilitator of the information diffusion than the other way around.

It is worth to point out here that even though we have demonstrated the robustness of the partition into communities we have to point out that the two communities analyzed in this section are not isolated from each other. As demonstrated earlier they are quite interlinked and each of them plays a considerable role for the functioning of the other.

5 Conclusion

In this paper we have taken the first attempt to investigate the complete blogosphere of a single country. The case has been that of Georgia that presents unique opportunities for collecting the information. The information about interlinkages among the 3272 Georgian blogs has been collected. Besides, where possible additional information about posting and commenting activities has been also collected. The dataset has been augmented by the post commentators that are not bloggers.

As it turned out Georgian blogosphere has a core-periphery structure with the core comprising around 65% of the whole network mass, as measured by the number of blogs. Potential speed of information diffusion in this network is quite high as the average shortest path length and the network diameter stand at low values. We have also detected two large communities in the network. They have been identified based on their linkage behavior. These communities differ in their posting and linkage behavior. There is also a difference in how long discussions the posts from these communities can trigger.

However, there is great deal more that can and has to be done especially to better understand the differences between the two communities. Firstly it would be interesting to analyze the commentators of posts. In particular it is interesting whether the participants in the discussions triggered by a blogger also belong to the same community. It is also important to understand how much of an inter-community discussion is each community able to generate. This approach would lead us to understand whether discussions in Georgian blogosphere are contributing to polarization of the public opinion or on the contrary are working towards public consensus.
Secondly, the content of posts can be collected. The analysis of collected content will allow a researcher to understand whether the separation between the two communities also extends to the topics they are blogging about. This will also allow for drawing the conclusions about whether the two communities are separated because of the difference in topics that they blog about or because of the differences in opinion.

Thirdly, the most intriguing continuation of the research would be to further utilize our blogging database and fill it with the new data on a regular basis. Continuing the exercise for few years (which becomes quite feasible as costs of information collection decrease substantially once we have already built the database) will give us the opportunity to research the evolution of Georgian blogosphere. With this data we can identify not only the evolution of the whole network or that of communities, but also the evolution of topics that Georgians are concerned about.
6 References


## Appendix

<table>
<thead>
<tr>
<th>rank</th>
<th>in-degree</th>
<th>out-degree</th>
<th>betweenness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sweet.g</td>
<td>ermonia.b</td>
<td>maisuradzenatia.b</td>
</tr>
<tr>
<td>2</td>
<td>dodka.g</td>
<td>mmiramax.b</td>
<td>litterator.w</td>
</tr>
<tr>
<td>3</td>
<td>tomushka.b</td>
<td>zurriuss.w</td>
<td>portxlebi.w</td>
</tr>
<tr>
<td>4</td>
<td>tiny.g</td>
<td>weromania.b</td>
<td>datomalania.b</td>
</tr>
<tr>
<td>5</td>
<td>embrioni.b</td>
<td>dgiurebi.b</td>
<td>ermonia.b</td>
</tr>
<tr>
<td>6</td>
<td>dgiuri.c</td>
<td>spacecrusader.w</td>
<td>kevana.w</td>
</tr>
<tr>
<td>7</td>
<td>kate9alukard.w</td>
<td>konnchetina.b</td>
<td>faceofware.b</td>
</tr>
<tr>
<td>8</td>
<td>bigcrow.w</td>
<td>busx.w</td>
<td>bigcrow.w</td>
</tr>
<tr>
<td>9</td>
<td>zurriuss.g</td>
<td>mecniavarkhar.w</td>
<td>burusi.w</td>
</tr>
<tr>
<td>10</td>
<td>iasamani.w</td>
<td>e2e4paiki.w</td>
<td>embrioni.b</td>
</tr>
</tbody>
</table>

Table 4: The list of top Georgian blogs by different characteristics.

Notes: “http://” is assumed in front of the name of every blog reported. Suffixes are abbreviated according to the following rule: .w = .wordpress.com, .b = .blogspot.com, .g = .ge and .c = .com.

<table>
<thead>
<tr>
<th>rank</th>
<th>community 1</th>
<th>community 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>kate9alukard.w</td>
<td>sweet.g</td>
</tr>
<tr>
<td>2</td>
<td>dodka.g</td>
<td>bigcrow.w</td>
</tr>
<tr>
<td>3</td>
<td>embrioni.b</td>
<td>iasamani.w</td>
</tr>
<tr>
<td>4</td>
<td>tiny.g</td>
<td>tomushka.b</td>
</tr>
<tr>
<td>5</td>
<td>zurriuss.g</td>
<td>norteri.b</td>
</tr>
</tbody>
</table>

Table 5: Community leaders, as ranked by within-community in-degree.

Notes: “http://” is assumed in front of the name of every blog reported. Suffixes are abbreviated according to the following rule: .w = .wordpress.com, .b = .blogspot.com, .g = .ge and .c = .com.