Interdisciplinary Description of Complex Systems 2(2), 126-135, 2004

TOWARDS QUANTITATIVE TOOLS FOR ANALYSING QUALITATIVE PROPERTIES OF VIRTUAL COMMUNITIES

Duje Bonacci*

Physical Chemistry department, Ruđer Bošković Institute
Zagreb, Croatia

Preliminary report *Received*: 10 May, 2004. *Accepted*: 3 November, 2004.

SUMMARY

During the last decade, the advance of Internet has enabled the emergence of previously nonexistent type of human social structures - virtual '*online*' communities. As compared to the traditional communities, online communities are distinguished by the drastic reduction of the requirement for the physical proximity and geographical clustering of their members. The primary cause of this shift away from 'physically concentrated' communities to dispersed virtual ones is new long distance communication tools that Internet has provided. Along with the increase in quantity of communication that the new technology brought about, it also strongly influenced its quality. The paper suggests two simple mathematical tools for analysing the 'soft' (qualitative) sociological internal properties of virtual communities. The suggested tools are applied and their utility discussed on the example of one such virtual community, Croatian NGO 'Society znanost.org'.

KEY WORDS

mailing list analysis, online communities, scale-free distribution, quantitative analysis of qualitative properties, social energy

CLASSIFICATION

PACS: 89.20.Hh

*Corresponding author, η: duje@znanost.org; +385 98 469761;
Physical chemistry department, Ruđer Bošković Institute, Bijenička 54, HR - 10000 Zagreb, Croatia

INTRODUCTION

VIRTUAL COMMUNITIES

Emergence of virtual communities

During the last decade, the advance of Internet has enabled creation of previously nonexistent type of human social groups - virtual 'online' communities [1]. As compared to the traditional geographically highly condensed communities, online communities are distinguished by the complete eradication of the requirement for the physical proximity and geographical clustering of their members. As the basic ingredient of community is the existence of reliable communication channels among its members, the primary causes of this shift away from 'physically concentrated' communities to dispersed virtual ones are the new long distance communication tools that Internet has provided.

Certainly, tools and services for long distance communication themselves are not novelty. Couriers and mail have existed for centuries, telegraph has been with us for almost two hundred years and during past century telephone and fax have joined this arsenal. However, the emergence of the community from the group of previously unrelated individuals requires the possibility of engagement of these individuals in a number of frequent and intensive formal and informal communication. 'Classic' communication tools mentioned earlier just could not support these requirements. Couriers and mail are relatively slow, as they both require physical transport of the message between geographically remote destinations and hence intensity of the information exchange is strongly compromised. Telegraph, telephone and fax have managed to drastically cut down on the message transmission time by using fast electromagnetic signals. They are the 'real time' communication tools and hence could possibly support the required intensity of the discussion. However, they also have two limitations. First, their extensive usage is relatively expensive for average consumer - and usually the price grows with geographical distance between users - and hence is mostly used 'in emergency' and only for important formal discussions. As such, they don't provide the possibility of frequent and informal communication. Second, they are 'one-to-one' means of communication and do not support multi-party discussions involving more than two participants at the same time. Hence they cannot convey the feeling of participation in real 'community' discussion to their users.

The affordability of home computers, the rapid spread of global computer network - the Internet - has cut down the price of real-time long-distance communication to just a small percent of its cost some ten years ago. Also, development of 'user friendly' communication software (web browsers, e-mail clients, internet telephony tools) has eliminated the need for acquiring any particular expert knowledge in order to be able to use these services on Internet. Hence, the new technology has opened the door to growth of new kind of physically dispersed global communities, as well as provided a new means for sustaining and reinforcing the 'classical' physically concentrated ones. As one of the greatest media theorists, Marshal McLuhan, would put it, our planet has indeed become a 'global village'. Further, not only did the technology make available new efficient 'broadband' communication channels and hence enabled the massive increase in the quantity of communication, but it also had far-reaching impact on the quality of our relation with fellow humans, physical environment we inhabit and even our own inner mental landscape [2].

Research into virtual communities

Social researchers have been interested in virtual Internet communities from their very beginning in the early 90s [3, 4]. This interest was supported by the fact that the very nature of the new communication tools made the communication itself amenable to quantification. The digital data format of the new online communication made it feasible to archive, retrieve, classify, sort, backtrack and analyse massive amounts of this communication in many different ways. During past century, scientists analysing quantitative mathematical aspects of networked structures - including human networks such as online communities - have discovered a surprising regularity that these seemingly completely random structures possess, commonly referred to as **Pareto's law** or **Zipf's law** [5 - 7]. Recently the deep relation between Pareto's law and dynamical rules governing the emergence and evolution of the various types of natural (e.g., ecosystems and crystals), social (e.g., companies and cities) and artificial (e.g., computer networks and power grids) networks has been established [8 - 10].

Research aim

This somewhat surprising encounter of strict mathematically expressible regularities and rather ‘soft’ world of social science research gives a hope that a precise mathematical toolkit is foreseeable which could be used to study the new sociological dimension of communication that emerged along with the virtual communities. The primary aim of the work presented in this paper was to present and discuss some rough-cut and simple, yet potentially useful quantitative tools for analyzing the quality of communication within the virtual communities.

ZNANOST.ORG SOCIETY

Brief history of znanost.org

Society znanost.org (in Croatian, *znanost* translates as *science*) is a Croatian non-government non-profit organization (NGO). It was initiated in early 2002 by the group of recently graduated physics students from Zagreb University. During past two years the membership has spread both to other natural sciences, as well as to social and technical ones. The mission of the **znanost.org** is to organise, promote and support activities that would help Croatia towards becoming the knowledge-based society. More information about the **Society znanost.org** and its activities can be found on the official web pages, http://www.znanost.org.

The main activities of the society are not directed towards amassing the membership, but rather towards fostering various projects aimed at promoting and disseminating knowledge as a tool of choice in Croatian society at large. Any interested party - whether member or non-member of **znanost.org** - that wants to start a project whose goals are in line with the overall mission of **znanost.org** can submit a project proposal to **znanost.org** Projects Board asking for the support. If the project is accepted, **znanost.org** provides the supporting infrastructure for the project - web space on society's server, financial administration services and network of contacts that society has developed with Croatian scientific and journalist community. Society, however, does not provide immediate financial support for the projects, as it has no sources of income of its own. Rather, it provides 'branding' that can help supported projects in finding commercial or charitable sponsorship.

Internal structure of znanost.org

The real structure of **Society znanost.org** is highly non-hierarchical. However, to be registered as an NGO in Croatia, the society has to have all the necessary formal structure elements required by the current Croatian law: President, Vice President, Secretary and Advisory Board with 3 members and Board of Members with at least 11 members. Due to strong inclination towards project-based activities, the society has also established a 3-member Projects Board which on the behalf of the society examines the submitted project proposals and declares its decision to the Board of Members.

Although in the first days of its existence great majority of the society's members were located in Croatian capital, Zagreb, in the following years this changed drastically. This change came about due to two reasons. First, some of the new members that were 'recruited' in the meantime were at the time already abroad. Also, significant number of initial members that are currently active in the society's activities has since left Croatia. All of these members that are currently abroad left the country exclusively for educational and professional purposes - they are now PhD students at various highly-rated scientific research institutions around the world, including Italy, Germany, Switzerland, United Kingdom and USA.

Communication within znanost.org

Almost from the very beginning - as it can be easily guessed from the very name - **znanost.org** was envisaged as a society with a strong underlying Internet backbone. It is quite questionable whether the society could function at all in its present, physically dislocated form if it did not have the whole range of practically free internet telecommunication tools at its disposal. Hence, during last year **znanost.org** has transformed into a fine example of literally ‘online’ community.

Through several major successful projects - just to mention the website (http://www.znanost.org) and press service of the First Croatian Science Festival - the society has raised sufficient funds to purchase its own web server. Thanks to several of the members' computer administration skills, this server has become the heart of the society's internal communication. Individual communication is done through direct e-mail correspondence, whereas majority of topical discussions are conducted on purpose-made mailing lists. As the server is completely administered by **znanost.org** members, the management of these e-communication resources within the society is very prompt and flexible. Some time ago, the free internet telephony tool, Skype (http://www.skype.com), has started getting some attention as a very useful prospective mean of immediate real-time communication. However, as it requires users to be online at the same time - which can be a bit of a problem having in mind their geographical dispersion - e-mail and mailing lists are still the tools of choice.

The oldest and probably the most influential mailing list of them all (within **znanost.org**) is the list *glavni@znanost.org*. This list is some sort of the society's virtual 'main square' - it serves as a general purpose bulletin board and chat room for the all the members of **znanost.org** administrative bodies. All the projects that at later stages develop their own separate mailing lists necessarily kick of from *glavni@znanost.org*. Also, voting procedure on all the major issues related to **znanost.org** activities is done through this list. Hence, the starting premise of this work was that if any of the society's general socio-dynamical characteristics is to be retrieved from the archived correspondence between its members, it must certainly be found precisely in the archive of *glavni@znanost.org*.

METHODS

SORTING THE POSTS

Two research methods employed in this paper are fully quantitative and actually belong to the realm of the natural rather than the social sciences. Both presented methods are fairly simple and, as will be demonstrated, they greatly differ with respect to their actual usefulness. It should be mentioned that the fully-fledged analysis could and should include much broader range of sociological approaches, primarily to strictly operationally define various elements of - and their respective contributions to - the concept of 'internal social energy' (i.e. the quantity and intensity of actual activities and internal social dynamics) of a particular virtual community. Indeed, some theoretical work on the closely related topic of social thermodynamics has already been done in discussing general social systems [11, 12]. However, this work was envisaged as the first - rather than the final - step towards new quantitative tools for analyzing the qualitative aspects of online communities. In that light, the presented analyses should be accepted as fully justified notwithstanding their obvious limitations.

Daily traffic analysis

The analysed data set comprised of 1634 posts sent by the members of **znanost.org** to the *glavni@znanost.org* mailing list between 2 March, 2003 and 4 November, 2003. The first quite obvious classification aimed at analysing the daily traffic on the list during this period and comparing it with the author's first-hand 'insider' knowledge of society's various activities. Put in other words, the aim was to find out whether there is a correlation between the 'virtual' activity of **znanost.org** society (reflected through the frequency of posts on the *glavni@znanost.org* mailing list) and its internal social energy. The main idea underlying this analysis was that if it was possible to establish some strong correlation between the two, then mailing list posting frequency could be generally utilised as a direct quantitative measure of the internal social energy of the virtual community.

The first set of data - the posting frequency - was extracted directly from the date stamps in the mentioned posts sent to the *glavni@znanost.org* mailing list. The respective set of data about the society's members' activities was based on the author's insider's background knowledge of the society and the range and dynamics of its activities.

Thread length distribution analysis

The second classification was concerned with threads that have emerged on the list. Thread develops in such a way that one member posts a message with some subject, other members reply to that mail with 'reply to' option in their e-mail clients automatically copying the subject line. This type of mailing list correspondence often results in a whole chain of replies and replies-to-replies etc. containing the same subject line. This chain shall be referred to as *thread*. Due to the complete freedom in the choice of initial subject line, thread analysis uniquely and sharply decomposes the mailing list archive into series of disjoint and easily discernible elements. The simplest way to classify threads is by their length, i.e. the total number of posts they contain. The second tool was the analysis of distribution of the number of threads vs. their length.

RESULTS AND DISCUSSION

DAILY TRAFFIC ANALYSIS

During the period from 2 March, 2003 to 4 November, 2003 (248 days) a total of 1634 posts were sent to the *glavni@znanost.org* mailing list by the list members. The resulting daily traffic on the list is displayed in Figure 1.

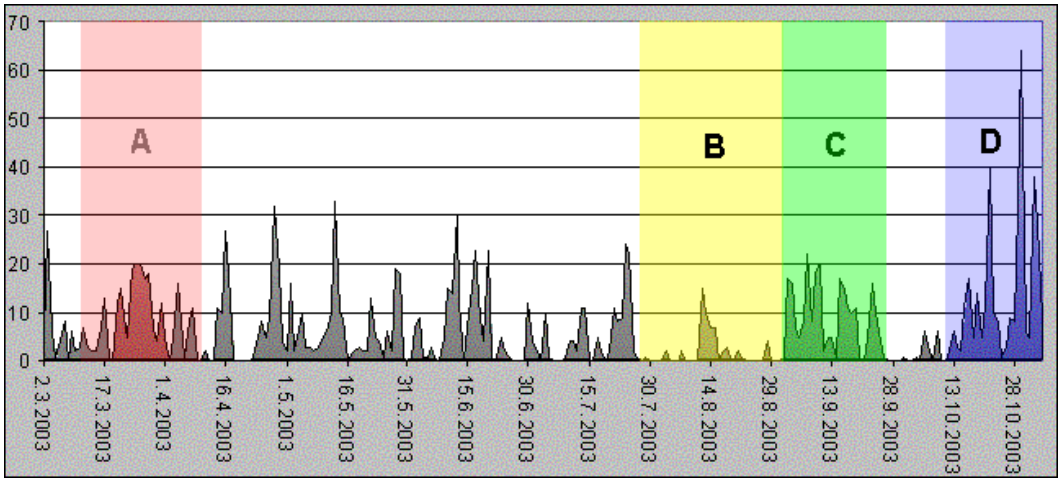

Figure 1. Daily traffic on *glavni@znanost.org* mailing list. Dates are on the x-axis, corresponding number of posts on the y-axis.

One of the characteristics of the society's internal dynamics which is quite notable from the daily traffic analysis is the 'burst mode' of its activities. The reason for this is simple to explain: due to the possibility of high intensity discussion that communication tools used provide, any problem or issue that arises within the society is dealt with very quickly. As all the members of **znanost.org** are volunteers and have their jobs to attend on daily basis, there is no regular daily 'background' correspondence on the *glavni@znanost.org* mailing list. The list is 'active' only occasionally, when there is some issue to be resolved. In this respect it would be interesting to analyse some mailing list which is tied to some more regular activity. One may argue that such a list would exhibit much smoother level of activity.

First Croatian Science Festival

One of the major events that **znanost.org** intensely worked on during the 2003 was First Croatian Science Festival. Festival was held during the week of 12-18 May, but as members of znanost.org were on the organising committee of the event, the activities related to it begun roughly about two and a half months earlier, in early March.

The first relatively extended period of constant activity on the daily traffic graph between mid March and early April corresponds to the various preparations for Science Festival. In Figure 1 it is designated with letter A. Rough content analysis of the posts shows that during this period several other very important things were discussed on the list. Along with the Science Festival related issues, during this period the society was highly engaged in acquiring the server as well as applying for membership with Croatian Research and Academic Network, CARNet. This conundrum of 'big issues' has produced a relatively steady daily inflow of posts to the mailing list, since most of these issues were related to some 'physical' activities

(contact the computer hardware stores, contact CARNet's personnel, etc.) that members had to conduct on daily basis and report back to the list with the results.

However, it is interesting to note that list activity does not show any major systematic increase from this period towards the Festival. One of the probable reasons for this is that members involved in organisation of the Festival diverted the traffic related to the Festival to the separate mailing lists. Hence, the increase in the activity due to the Festival cannot be noted on the 'general purpose' *glavni@znanost.org* mailing list.

Summer break

Summer break is the second reasonably distinguishable period. On Figure 1 it is designated with letter B. It is characterised by the drastic decline in list activity between late July and late August. However, due to the difference in dates when various members went on their summer break - some in early July, others in late July - there is a sharp peak in list activity halfway through this period when they all have briefly 'met' again on the list.

End of the summer break

The second period of relatively sustained activity of the list, somewhat expectedly, started immediately after the summer break in early September. In Figure 1 it is designated with letter C. Rough content analysis of the posts shows that through this period several issues were 'finished off' that were started just before or during the summer break or in the mid August half way through the summer break.

Beginning of the academic year

The last distinguishable span on the daily traffic graph is the one at the end of the analysed period. On Figure 1 it is designated with letter D. It is interesting as it includes several highest peaks of the activity of the list during the whole analysed period. Content analysis shows that these peaks correspond to discussions on some major organizational issues (formulation and discussion of society's projects regulations bill and society's finances regulations bill) as well as the start of the engagement with several new projects.

To conclude: using the insider background knowledge of the **znanost.org** virtual community, hints of correlation between the mailing list activity and society's internal social energy can be glanced. However, it would be a gross overstatement to conclude that using this simple approach allows the full and precise quantification of that correlation. In other words, without the insider background knowledge, applying just simple posting frequency analysis to the other similar virtual communities seemingly does not enable one to firmly conclude on the precise periods of the community's actual activities, or any qualitative features of these activities. It would be a matter of further research to inquire into whether a more sophisticated analysis tool could be developed by somehow 'upgrading' the posting frequency analysis approach. Nevertheless, this method can readily be utilised as a neat 'bookkeeping' tool for analysing, summarizing and presentation of particular virtual community's activity in a given period.

THREAD LENGTH DISTRIBUTION ANALYSIS

The second analysis the data was subjected to is the thread length distribution analysis. Namely it could be argued that thread length distribution correlates in some way with internal structure and actual activities' dynamics of the virtual community composed of all the members of particular mailing list. So, for example, one of the signatures of the existence of virtual community with high value of internal social energy would be that there are some long threads, as long threads necessarily result from a prolonged and coherent exchange of information. To compare the lists with very different number of members, obtained results should be suitably normalised, e.g. by considering not total but relative number of posts per thread, obtained by dividing each thread length with the total number of list members.

Certainly, to prove such an assertion comparative analysis should be conducted using data from whole number of mailing lists. However, from the point of view of introductory research such as this one is, it suffices to focus the attention on a very simple task: to see whether data on thread length distribution from this single virtual community's mailing list provides some firm quantitative parameters. In further research, such parameters could be used as a quantitative measure of internal social energy of the virtual community.

The data decomposed in total of 446 threads with lengths ranging from 1 to 34. The plot of thread length distribution is given in Figure 2.

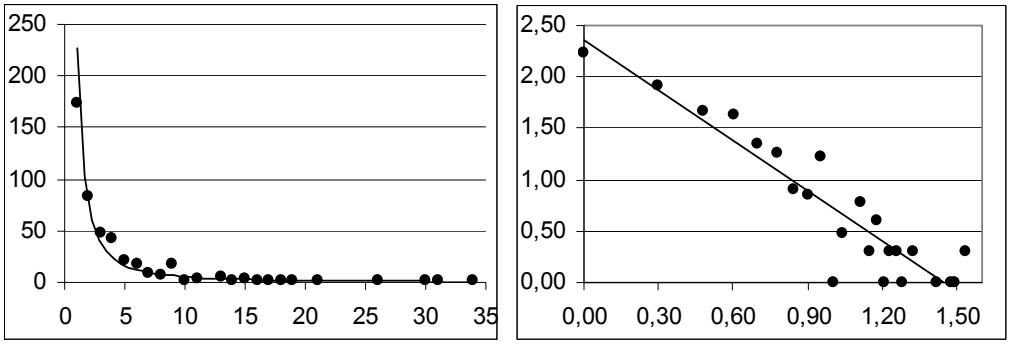

Figure 2. Thread length distribution for *glavni@znanost.org* mailing list. On the left graph, thread length n is on x-axis, number of threads $N(n)$ is on y-axis. On the right the same data are reproduced, but as a log-log plot of thread length distribution. The solid lines are fit to the data using least squares method.

A very impressive characteristic of the obtained distribution is that it is not random but rather highly ordered. Indeed, when data are fitted using the least squares method to the power-law curve of the form:

$$N(n) = A \cdot n^{B}, \tag{1}$$

a high correlation coefficient of 0,86 is obtained, with parameters $A = 227$ and $B = 1{,}6$. This trend-line is also displayed on left graph in Figure 2. On the right graph in Figure 3 the same data is reproduced in the log-log scale from which the fit of data to the trend-line (which in this plot becomes a straight line) is even more evident. Hence, the thread length distribution on the *glavni@znanost.org* mailing list follows the power law.

This finding suggests that thread length analysis could possibly yield some firm quantitative marker of the community's internal social energy. It could be suggested that this power-law distribution relates to the fact that members of the

list actually form a very strong self-organised network. Indeed, it has been shown [9, 10] that principles of self-organisation in social networks do lead to (scale free) power law distributions. However, further research should be conducted in order to fully explore this assertion.

CONCLUSIONS AND FURTHER RESEARCH

As it was stated already at the very end of the Section 3.1., the first of the method of daily traffic analysis is on its own of a very limited value as a tool for analysing the qualitative elements of the virtual community's inner dynamics. Nevertheless, the possibility still remains that if it used in conjunction with some more elaborate sociological methods it can provide some quantitative reinforcement for the conclusions based on pure qualitative observations. In this respect, although this paper offers no conclusive findings, further research on the topic could yield some interesting and useful results.

On the other hand, even in the rough form presented here, the method of thread length analysis seems to offer a very promising quantitative probe of virtual communities' internal social energy. Certainly, further and more elaborate research is required to develop proper benchmarks that would allow the comparison and classification of different virtual communities. First step that should be done along these lines would be to extend the analysis conducted in this work to other mailing lists in order to determine whether some reliable quantitative 'markers' (e.g. like parameters *A* and *B* of the power law distribution (1)) for various qualitative parameters of virtual communities' inner dynamics and sociology can be established. Such markers could be useful as a tool for quickly selecting particular virtual communities on which more extensive (and laborious) sociological, ethnographic or other more in-depth qualitative analysis is to be conducted.

ACKNOWLEDGMENTS

I would like to thank my friend and colleague Damir Kovačić for providing me with the data for this research. I would also like to thank all of my other 'comrades' from **Society znanost.org** for highly creative and inspiring online environment they provide within our common efforts in bringing knowledge higher on the agenda of contemporary Croatian society.

REFERENCES

[1] Boetcher, S.; Duggan, H. and White, N.: *What is a Virtual Community and Why Would You Ever Need One?*
Full Circle Associates, 2002,
http://www.fullcirc.com/community/communitywhatwhy.htm,

[2] Press, L.: *McLuhan meets the net*.
Communications of the ACM **38**(7) 15-20, 1995,
http://som.csudh.edu/cis/lpress/articles/macl.htm,

[3] Beckers, D.: *Research into virtual communities: an empirical approach.* PDC '98 / CSCW '98 Workshop 'Designing Across Borders: The Community Design of Community Networks', Seattle, 1998,
http://www.swi.psy.uva.nl/usr/beckers/publications/seattle.html,

[4] Silver, D.: *Looking Backwards, Looking Forward: Cyberculture Studies 1990-2000.*
Gauntlett, D., ed.: *Web studies: Rewiring Media Studies for the Digital Age*, Oxford University Press, 19-30, 2000,
http://www.com.washington.edu/rccs/intro.asp,

[5] Dalziel, P.; Higgins, J.: *Pareto, Parsons and the boundary between economics and sociology*.
The 15th HETSA Conference, Armidale, 2002,
http://nzae.org.nz/files/%2365-DALZIEL-HIGGINS.PDF,

[6] Dugan, M.: *Pareto's Law, a management principle/technique*.
http://home.alltel.net/mikeric/Misc/Pareto.htm,

[7] Knudsen, T.: *Zipf's Law for Cities and Beyond: The Case of Denmark*.
American Journal of Economics and Sociology **61**(1), 101-121, 2001,
http://www.findarticles.com/cf_dls/m0254/1_60/74643763/p1/article.jhtml,

[8] Matlis, J.: *Scale-Free Networks*.
Computerworld knowledge center – networking, November 2004,
http://www.computerworld.com/networkingtopics/networking/story/0,10801,75539,00.html,

[9] Barabási, A.L.; Reka, A.: *Emergence of scaling in random networks*.
Science **286**, 509-512, 1999,
http://www.nd.edu/~networks/Papers/science.pdf,

[10] Newman, M.E.J.; Watts, D.J. and Strogatz, S.H.: *Random graph models of social networks*.
Proceedings of the National Academy of Sciences of the United States of America – Physical Sciences **99**(Suppl. 1), 2566-2572, 2002,
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=128577,

[11] Stepanić, J.; Štefančić, H; Žebec, M.S. and Peračković, K.: *Approach to a Quantitative Description of Social Systems Based on Thermodynamic Formalism.*
Entropy **2**, 98-105, 2000,
http://www.mdpi.org/entropy/papers/e2030098.pdf,

[12] Stepanić, J.: *Social Equivalent of Free Energy.*
INDECS **2**(1), 53-60, 2004,
http://indecs.znanost.org/2004/indecs2004-pp53-60.pdf.

DOPRINOS KVANTITATIVNIM ALATIMA ZA IZUČAVANJE KAKVOSNIH ZNAČAJKI VIRTUALNIH ZAJEDNICA

Duje Bonacci

Institut Ruđer Bošković
Zagreb, Hrvatska

SAŽETAK

Tijekom posljednjeg desetljeća, razvoj je Interneta omogućio nastajanje ranije nepostojećih vrsta ljudskih društvenih građevina - virtualnih '*mrežnih*' zajednica. U usporedbi s tradicionalnim zajednicama, mrežne se zajednice odlikuju gotovo potpunom neosjetljivošću na fizičku udaljenost i geografsku usredotočenost njihovih članova. Prvenstveni uzročnici ovog pomaka od 'fizički usredotočenih' zajednica prema potpuno raspršenim virtualnima jesu nove na Internetu zasnovane telekomunikacijske tehnologije. Usporedo s omogućavanjem povećanja obima komunikacije, nove su tehnologije također ostavile bitan i trajan pečat i na njezinu kvalitetu. U ovom radu su predložene dva matematička alata za izučavanje 'mekih' (kakvosnih) sociooloških značajki virtualnih zajednica. Ujedno, izložen je i primjer primjene ovih alata na jednoj takvoj virtualnoj zajednici, hrvatskoj nevladinoj udruzi 'Društvo znanost.org'.

KLJUČNE RIJEČI

analiza liste e-pošte, mrežne zajednice, bezljestvična raspodjela, kvantitativno izučavanje kakvosnih svojstava