

THE NETSCAN
PROJECT HELPS
ONLINE PARTICIPANTS
FORM COOPERATIVE
RELATIONSHIPS BY
OFFERING A BETTER
SENSE OF THE OTHER
PLAYERS INVOLVED.

TOOLS FOR NAVIGATING LARGE SOCIAL CYBERSPACES

MARC SMITH

USENET, THE GLOBALLY DISTRIBUTED DATABASE OF CONVERSATION AND OTHER MATERIAL, IS A POWERFUL EXAMPLE OF ANARCHIC SOCIAL ORGANIZATION. NO ONE IS IN CHARGE OF USENET, NO CENTRAL AUTHORITY CONTROLS ITS BORDERS OR CONTENT IN THE WAY SO MANY COMMERCIAL SERVICES ARE RULED BY SINGLE COMPANIES OR GROUPS. IN 2000, AT LEAST 8 MILLION UNIQUE PARTICIPANTS CREATED MORE THAN 150 MILLION MESSAGES UNEVENLY DISTRIBUTED OVER 50,000 OR MORE NEWSGROUPS DEVOTED TO EVERY TOPIC OF POSSIBLE HUMAN INTEREST, FROM RESELLING TAIWANESE HOUSEHOLD GOODS, TO DEBATING RELIGION, TO TRADING SOFTWARE. LIKE MANY RELATED CONVERSATIONAL MEDIA, INCLUDING EMAIL LISTS AND WEB DISCUSSION BOARDS, THE PROBLEM IS OFTEN NOT FINDING OTHERS WHO SHARE YOUR INTERESTS. INSTEAD, THE CHALLENGE IS DEALING WITH TOO MUCH CONTENT OF MIXED VALUE.

Ideally, Usenet members would make efficient use of bandwidth, participating actively but judiciously in newsgroups, ensuring their comments are posted only to relevant newsgroups, and abiding by the local norms and culture that govern decorum. Everyone is better

off if all behave in such a manner, but there is the temptation to free-ride on the efforts of others. Thus, some participants post articles that are unnecessarily long, or lurk rather than contribute to the give-and-take that is an essential feature

of any newsgroup, or post off-topic articles, or violate the local rules of decorum. The more people free-ride, the more difficult it is to produce useful information and interaction. In the language of the Usenet, the signal-to-noise ratio deteriorates. Newsgroups silt up with commercial messages, fraud, and gibberish. The challenge becomes a question of how a group of individuals can “organize and govern themselves to obtain collective benefits in situations where the temptations to free-ride and to break commitments are substantial” [5].

A key finding of collective action studies shows that mutual awareness of other participants’ histories and relationships is critical to a cooperative outcome. The challenges of cooperation are heightened further when people are able to draw from a resource without contribu-

LISA HANNEY

tion and when contributors are difficult to identify. It would make many of the social processes that bind collective projects together more effective if participants in social cyberspaces could easily create and access histories of one another and the spaces they interact within [2]. But the tools available to interact with these social spaces compound the problem.

Interfaces, like email and news browsers, that provide access to social cyberspaces such as discussion boards, email lists, and chat rooms, present limited, if any, information about the social context of the interactions they host. Basic social cues about the size and nature of groups are missing, making discovery, navigation, and self-regulation an increasing challenge as the size and scope of these spaces expand. While people can eventually develop a refined sense of the rhythms, leaders, and fools in a particular social cyberspace, the information does not come easily or easily transfer to other spaces. With little sense of the presence of other people, individuals have a difficult time forming cooperative relationships.

as the number of days on which each author contributed a message to the newsgroup, or the fraction of each author's messages that were replies to existing messages versus initial turns or postings. Combined with information about the structure and development of threads, the chains of turns and responses created in these spaces, these metrics can be used to extract content of likely value out of large, active social cyberspaces.

Various aspects of people and conversations can be used as guides through the welter of content depending on the kind of task pursued. Such measures can be used to separate active groups from dormant or dead ones and to distinguish discussion newsgroups from job listings and other non-interactive newsgroup types. The existing process of newsgroup discovery is hit-or-miss since as much as two-thirds of all newsgroups are desolate or mostly inactive. More than 100,000 newsgroups exist but only about 50,000 receive more than a single message a day. Only 20,000 receive more than one a day. Most news browsers offer no indication of the population of

A key finding of collective action studies shows a mutual awareness of other participants' histories and relationships is CRITICAL TO A COOPERATIVE OUTCOME.

The goal of the Netscan project at Microsoft Research is to generate social accounting metrics of a range of dimensions of social cyberspaces in general and the public Usenet feed in particular. Netscan is a publicly available Web service located at netscan.research.microsoft.com. Social accounting metrics are measures of the social dimensions of online spaces, like the size of newsgroups in terms of messages and participants along with the histories of the participating authors in each, such as how long they have been active in the group, in what other newsgroups they participate, to what threads of conversation they contribute, and which other participants they most often engage in discussion [4]. Standard news browsers that present information about the messages themselves—their posting date, their subject, how many lines of text they contain—force users to pay more attention to the structure of the medium than the qualities of the participants who would more naturally draw the user's focus.

A news browser would serve users far better, for example, if it allowed users to sort and search newsgroup content by such salient behavioral attributes

newsgroups or their general levels of activity over time. As a result, finding an active newsgroup on an appropriate topic can be a chore.

To address this, the Netscan Web interface provides a newsgroup search engine that takes a string as input and returns a list of all the newsgroups containing that string in their name (see Figure 1). Most news browsers provide a similar feature but return a list of newsgroups without any indication of the social properties of the space. The Netscan interface reports the number of messages and authors that are present in each newsgroup returned by the query. These measures help distinguish active from inactive newsgroups and broadcast newsgroups (which have many messages but few authors) and more diverse newsgroups with many contributors. The presence of a persistent group of participants is indicated in a report of "returnees," authors who have come back to the newsgroup on a frequent basis. They are the newsgroup's regulars and the size of the population of regulars is an indicator of the maturity and stability of the space. Newsgroups in which few people are retained month to month may have limited value,

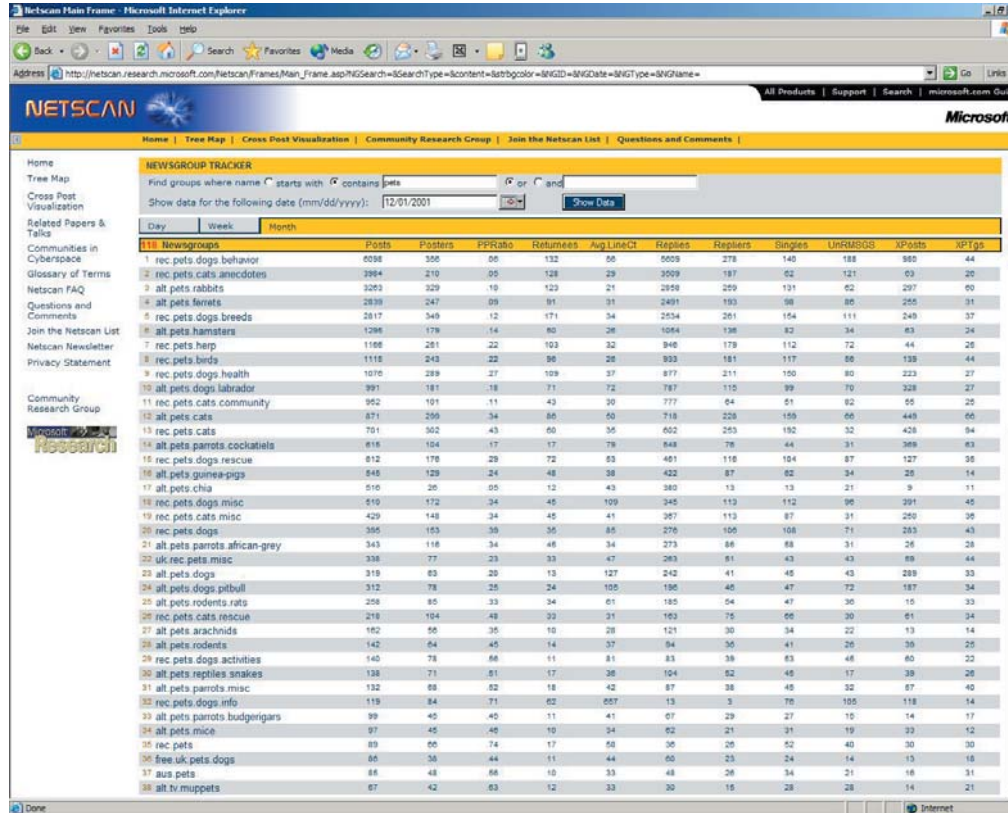


Figure 1. The Netscan home page reporting the result of a query for “pets”-related newsgroups

while those with large groups of people who have several months or more tenure in the newsgroup may indicate a more active group with a potential for a collective memory and the emergence of acknowledged leaders or authorities. Of course, many people could return frequently to a newsgroup and not actually respond to one another. The count of the number of replies and repliers is also reported to distinguish conversational from nonconversational newsgroups.

Newsgroups like *alt.binaries.sounds.mp3* and many other *binaries* newsgroups where software and multimedia objects are exchanged are the largest in terms of total messages, receiving nearly a million messages in 2000. But these newsgroups attract relatively few authors and minuscule amounts of replies. The newsgroups with the largest numbers of unique authors are devoted to marketplaces for used goods and discussions of technology. The largest newsgroup by population, *tw.bbs.forsale*, attracted 53,000 unique participants in 2000. Despite their population size, these newsgroups are actually remarkably nonresponsive, only about 2% of the messages in these newsgroups are replies. In contrast, while they have a smaller population, the newsgroups with the highest rates of reply are devoted to topics like *alt.atheism* and *alt.fan.rush-limbaugh*, controversial topics devoted to discussion and debate with rates of reply as high as 96%.

Defining the signature indicators for a variety of newsgroup types may help us construct interfaces that

enhance the discovery of interesting discussions and improve users’ awareness of the dynamics and notable members of newsgroups. Successful social processes rely upon a group’s awareness of the various levels of contribution that members provide to group projects, the size of the group, and its relationship to other groups. Introducing such information into social cyberspaces may help orient participants and support self-regulation and boundary maintenance activity that could improve content quality and user satisfaction [7]. To explore whether our approach is effective, the Netscan project will continue to evolve to the point where we hope to encourage a large population to use the service to browse and contribute content to newsgroups. Through field studies we hope to be able to show significant difference in the patterns of contribution in newsgroups predominantly accessed through socially enhanced interfaces.

The existing Netscan Web service demonstrates two possible approaches, a *thread tracker* and an *author tracker* that sift content based on social properties of online conversational spaces and their participants (see Figure 2). The thread tracker report displays the 40 largest threads in each newsgroup. Threads are mostly short structures; about 50% of all threads in Usenet in 2000 were made up of only two posts. The remaining threads range from 3 to nearly 20,000 messages, although only 0.03% of all threads in Usenet in 2000 were larger than 100 messages. For many newsgroups the largest threads are made up of 20 or 30 messages. In some cases newsgroups are touched by a few very large threads that are cross-posted or shared with 20 or 30 different newsgroups.

The thread tracker report often captures topics of broad interest to the community and rarely displays subjects that suggest spam or even off-topic subjects.

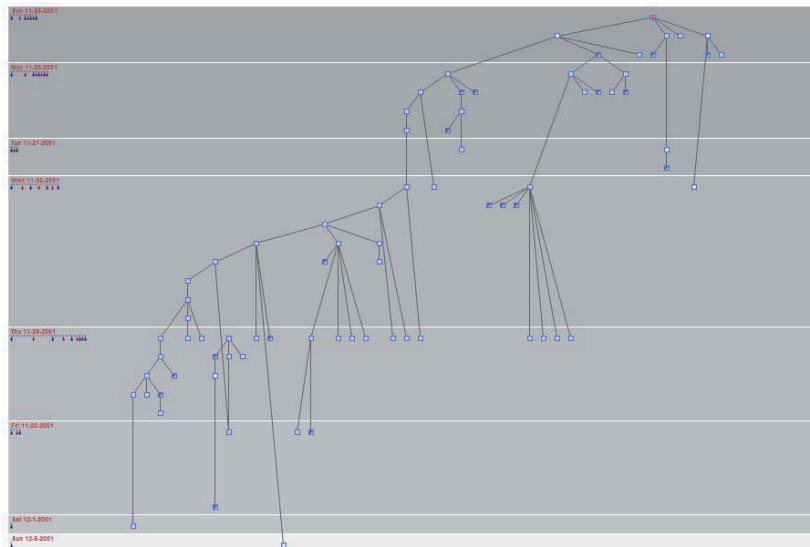


Figure 2. Thread tracker visualization of a large discussion thread highlights the structure and temporal development of threads.

The visualization component attempts to highlight important structural, temporal, and social dimensions of the conversation [1, 6]. Each message is represented as an icon shaded to indicate the role the message's author plays in the thread. The initiator of the thread is indicated with red-dotted icon, authors who posted only once in the thread are indicated with a white dot, the most active authors are represented with a half-shaded icon, and everyone else is simply blue. Each icon displays a tool tip containing the name of the author and their total messages in this thread on mouse-over. Message icons are displayed on top of gray stripes that indicate the calendar days the thread is active. Along the edge of the display is a histogram that indicates how many messages were posted on each day and how many authors posted those messages.

While the current Netscan report simply orders threads by the raw number of messages it contains, more sophisticated measures could take into consideration properties of the conversation like the speed the thread developed and its overall life span, the number of different participants and their patterns of contribution to the thread, and the pattern of cross-posting. If interfaces supported sorting content along these dimensions it will become much easier to sift value from even very noisy social cyberspaces.

The author tracker addresses a clear limitation of the thread tracker report: some small threads have high value. How can objective behavioral measures be used to highlight value in the larger population of smaller-sized threads? One approach is to find a measurable aspect of each author's behavior that is a likely indicator of his or her value. Our first approach is to


highlight the authors with the longest tenure in the newsgroup. The number of different days a person comes back to a newsgroup can be thought of as what the economists call a "costly signal." These are claims to status that are difficult to forge because they are difficult to create, for example, a sure sign of wealth is to spend a great deal of money, a sure sign of strength is to lift heavy objects. In newsgroups, a long history of contribution is a difficult thing to create immediately and is possible to identify mechanically. As a result the count of days active in each newsgroup quickly identifies the people with the most commitment to the newsgroup.

Interestingly, the most frequently active people are not the same as those who post the most. While high levels of messages (particularly initial turns) are often indicators of spam, the most frequently active posters (those who post on the greatest number of different days) share a common tendency to predominantly reply to others instead of initiating new threads. Furthermore, authors who were active over long periods of time were predominantly contributing to shorter threads than those reported in the thread tracker report. But are they the most valuable threads? Is the content of these author's threads more reliable or useful than any randomly selected thread?

To address the value of these metrics for selecting content from newsgroups several of us performed a study to evaluate their relationship with subjective evaluations of content [3]. In a recent study we recruited 22 regular users of newsgroups to come to Microsoft to read their favorite newsgroups and tell us what they liked and disliked about news browsers and newsgroups. In the course of that study our subjects created detailed evaluations of 309 authors from a wide range of newsgroups. These evaluations were correlated with a collection of metrics that described each author and the threads in which they participated. We found high levels of correlation between our subjects rating authors highly with high rates of those authors returning frequently to the newsgroup, contributing to high numbers of different threads that were mostly modest in size, and low numbers of other newsgroups visited.

Our interviews with heavy newsgroup users confirm our sense that the author tracker generally identifies authors they recognize as valuable. A direction for future work is to look for evidence of false posi-

tives, authors who are frequently active but are otherwise disruptive. One potential indicator is the notion of *thread domination*, the average proportion of each thread that a particular author creates. Authors who are routinely the dominant author in every thread they participate in show initial indications of having lower value than those who participate widely but with a certain restraint.

Social accounting metrics show promise as the basis for improved interfaces to social cyberspaces. In many ways these measures can act as a sociological application of the Heisenberg Uncertainty Principle in which measurement causes a change in the things measured. In this case social accounting metrics highlight a range of behaviors and give participants awareness of aggregate activity and individual contributions. The result might be persistent online conversations more resistant to disruption than existing public social cyberspaces that are frequently overrun by a small group of disruptive participants. Future research will explore the comparative performance of newsgroups and other conversational media with different levels of social accounting feedback. 

REFERENCES

1. Donath, J., Karahalios, K., and Viegas, F. Visualizing conversation. In *Proceedings of the Hawaii International Conference on System Sciences*. (Jan. 1999).
2. Erickson, T., Smith, D.N., Kellogg, W.A., Laff, M.R., Richards, J.T., and Bradner, E. Socially translucent systems: Social proxies, persistent conversation, and the design of Babble. In *Proceedings of CHI 99: Human Factors in Computing Systems*. ACM Press, New York, 1999, 72–79.
3. Fiore, A., Teirnan, S.L., and Smith, M. Observed behavior and perceived value of authors in Usenet newsgroups: Bridging the gap. In *Proceedings of CHI 02* (Minneapolis, Apr. 20–25). ACM Press, New York.
4. Hill, W., and Hollan, J. History enriched data objects: Prototypes and policy issues. *The Information Society* 10.
5. Ostrom, E. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, New York, 1990.
6. Sack, W. Discourse diagrams: Interface design for very large-scale conversations. In *Proceedings of the Hawaii International Conference on System Sciences: Persistent Conversations Track*. (Jan. 2000).
7. Whittaker, S., Terveen, L., Hill, W., and Cherny, L. The dynamics of mass interaction. In *Proceedings of CSCW '98*. ACM Press, New York.

MARC SMITH (masmith@microsoft.com) is a sociologist at Microsoft Research, Redmond, WA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

© 2002 ACM 0002-0782/02/0400 \$5.00

Call For Nominations

The ACM Transactions on Graphics (ToG) is seeking nominations for the position of Editor-in-Chief. This volunteer position is for a three-year term, and is appointed by the ACM Publications Board with the advice of ACM SIGGRAPH.

The Editor-in-Chief is responsible for maintaining the highest editorial quality, for setting technical direction of the papers published in ToG, and for maintaining a reasonable pipeline of articles for publication. He/she has final say on acceptance of papers, size of the Editorial Board, and appointment of Associate Editors. The Editor-in-Chief is expected to adhere to the commitments expressed in the ACM Publishing Rights and Responsibilities Policy (<http://www.acm.org/pubs/rights.html>). The Editor-in-Chief works closely with a number of ACM Volunteer units (Publications Board, SIGGRAPH, other Journal Editors) as well as with the Publications Department staff at ACM Headquarters.

Nominations should include a c.v. of the nominee, as well as a brief statement of why the nominee should be considered. Self-nominations are acceptable. Deadline for submitting nominations is April 30.

Please send all nominations to:

Steve Cunningham [rsc@eos.csustan.edu]
Chair, for the Nominating Committee

David Ebert, Purdue University
Francois Sillion, INRIA
Richard T. Snodgrass, University of Arizona
Stephen Spencer, University of Washington