RC24708 (W0812-066) December 11, 2008
Computer Science

IBM Research Report

Lets Jam: Drawing from the Wisdom of Your Employees

Claudia Perlich, Yan Liu, Rick Lawrence, Wojciech Grye, Mary Helander, Chandan Reddy, Saharon Rosset
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Lets JAM: Drawing from the wisdom of your employees

Claudia Perlich, Yan Liu, Rick Lawrence,

Wojciech Gryc, Mary Helander,Chandan Reddy, Saharon Rosset

IBM T.J. Watson Research Center

P. O. Box 218

Yorktown Heights, NY 10598

Abstract

Every work environment gives rise to social and work-related interactions between employees as well as employees and customers. These interactions create a socially interlinked work environment that affects the motivation and perceived satisfaction of employees and thereby performance. It also hosts a wealth of intrinsic business knowledge and potentially competitive advantage. In addition, recent advances in technology have enabled new means of interaction through instant messaging, blogs and sometimes open discussion forums. In 2006, IBM hosted the Innovation Jam — a moderated online discussion among IBM worldwide employees and external contributors — with the objective of identifying innovative and promising "Big Ideas". We describe the data from the Jam and investigate several analytical approaches to address the challenge of understanding "how innovation happens". Specifically, we examine whether it is possible to identify characteristics of such discussions and the implicit social network that are more likely to lead to innovative ideas. We demonstrate the social network structure of data and its time dependence, and discuss the findings of both supervised and unsupervised data analytics.

1 Introduction

The importance of social aspects of the work environment and its impact on employee productivity and satisfaction is well recognized. An aspect that has received less attention is the intrinsic value of the social employee network as well as the implicit but hidden business knowledge and wisdom of the 'crowd of employees'. With the demands of commoditization and globalization, businesses have an increasing incentive to tap into these hidden resources of their entire work force. Companies are focusing extensively on innovation as a significant driver of the new ideas necessary to remain competitive in this evolving business climate and turn more and more often to their own employee-base for insights and suggestions.

The broader issue is how does a company foster innovation, and specifically how do we identify, extend, and capitalize on the new ideas that are created in the workforce within the company?

With the wide use of worldwide web, people are provided a much more convenient and quick means for communication so that much larger and richer "virtual" social networks are formed, such as "MySpace", "Facebook" and "LinkedIn". Can such 'virtual' networks provide business value? Very often, those Web 2.0 sites are filled with gossip or, even worse, registered user names but no site visitors. Still, there are some prime examples where social networks really are driving business innovation.

For instance, consider a few recent examples of online forums within corporate environemts [16] where people can discuss topics of interest online at any time and any place. Firms increasingly try to take advantage of such virtual networks and to capture innovative ideas. One of the first efforts was IBM' Innovation Jam in 2006 [11, 17] followed soon by Dell's

IdeaStorm in 2007 (http://www.dellideastorm.com/) and many others thereafter. In both cases employees and external participants (customers and business partners) are encouraged to share their ideas around topics of broad interest. Analysis of the information collected in such forums requires a number of advanced data processing steps including extraction of dominant, recurring themes and ultimately characterization of the degree of innovation represented by the various discussion threads created in the forum.

We have obtained the complete set of comments generated by the participants of IBM's 2006 Innovation Jam. The Jam consisted of two successive phases, followed by a selection of highly promising ideas based on these discussions. This multi-stage format provides a rich set of data for investigation of how ideas evolve via moderated discussion. Of particular interest is whether we can detect characteristics of those discussion threads that can be linked to an idea ultimately being selected as a promising initiative. To the extent that selected ideas reflect some indication of "innovation," we have some basis for examining which aspects of a discussion thread may lead to innovation. This paper summarizes our efforts to characterize successful threads in terms of features drawn from both the thread content as well as social network analysis drawing from the tremendous work on the social network study over the past century [22, 2, 15].

The paper is organized as follows. In Section 2, we discuss some metrics that have been used to characterize the structure of social networks formed via other kinds of discussion groups. Section 3 describes the specifics of the IBM Innovation Jam and the collected data. Section 4 summarizes some key aspects of the dynamics of the Jam interactions. Finally, Sections 5 and 6 describe respectively the unsupervised and supervised learning approaches we have applied to this data.

2 Related Work

Understanding the nature and process of innovation has always been one of the major topics of interest, both from a practical point of business management and a theoretical point of innovation research [14]. The important connection between social interaction and innovation was recently revisited from a practical view in 'Social Networking: The Essence of Innovation' by Jay Liebowitz [18] and a scientific view in 'Research Issues in Social computing' by Parameswaran and Whinston [20]. Earlier work by Burt [6] points at specific network properties such as structural holes and their relationship to innovation and the rise of good ideas.

Olpert and Damodaran [19] point in the framework of e-Governance at yet another effect of broad engagement of participants in the innovation process: increased success in development and technology adoption. The importance of social position was also noted in reference to personal and team success by Ashworth and Carley [3].

Online forums like the IBM Innovation Jam or Dell's IdeaStorm are threaded discussions, whereby participants create topics and explicitly reply to each other using a "reply to" button. As such, discussions are hierarchical, with a clear flow of messages from the initial parent post to subsequent ideas and thoughts. Unlike more unstructured discussions, the person's response is then only associated with the message he or she chose to reply to. Through studies focusing on usenet groups and online forums, a great deal is known about the types of social structures that may develop within such communities. Indeed, it has been observed that depending on discussion topics, group membership and other features of a usenet group or forum, differing social structures may develop.

For instance, Turner et al. [23] show that newsgroups vary a great deal based on the amount of time individuals spend on a single visit, and the longevity of individuals' involvement in the groups themselves. The authors specifically break down the social roles of participants into types like the "questioner", "answer person", "spammer", and "conversationalist", among others. Depending on the participants, such groups can then develop different forms of participation and norms. For example, a questioner may post once in a group and never again, while an answer person may respond to a large number of questions, but never participate in continuing dialogues. A group of conversationalists can be characterized by a large number of posts and replies to each other.

While the social network analysis of Innovation Jam participants is beyond the scope of the paper, it is important to note that the social structure of the Jam has an effect on its outcomes. Newsgroups or forums can develop their own internal social structures. Groups can also be seen as those that develop unique linguistic or social traits [9]. For example, groups may begin to use new words or acronyms, lending support to the idea that predicting convergence in discussions or brainstorms is possible.

As described in the following section, the Innovation Jam format has been used extensively within IBM. The concept has also been extended to external events like Habitat Jam, at the World Urban Forum (www.wuf3-fum3.ca) in 2006. The brainstorming format within a limited time frame is recognized as a useful approach to dispersed and large-scale collaboration. Thus, it is useful to analyze the discussions, and explore how people reached the resultant "big deas" of the Jam.

3 Innovation Jam Background

In 2001, IBM introduced the Jam concept through a social computing experiment to engage large portions of its global workforce in a web-based, moderated brainstorming exercise over three days [10]. What became known as the "World Jam" was eventually followed by six additional internal, corporate-wide Jams, drawing employees into discussions about everything from management to company values. In early 2006, IBM announced that it would again use the Jam concept for an eighth time - this time, for facilitating innovation among the masses, and also including participants from external organizations and IBM employee family members.

Key to the design of the Jam's large scale collaborative brainstorming methodology was the identification of seed areas. Before the event launch, teams were formed to brainstorm general areas and to discuss the design and implementation details. Four general areas, called "*Forums*," were identified:

- **Going Places** - Transforming travel, transportation, recreation and entertainment
- **Finance & Commerce** - The changing nature of global business and commerce
- **Staying Healthy** - The science and business of well-being
- **A Better Planet** - Balancing economic and environmental priorities

Factors that determined the selection of seed areas included the opinions of IBM technical leaders, the technical relevance to IBM's business strategies, as well as the overall relevance to general societal and global economic challenges.

3.1 The Innovation Jam Process

IBM's Innovation Jam was designed to take part over two phases. Phase 1 took place July 24-27, 2006 and primarily focused on idea creation and development. Unlike previous IBM Jams where preparation was not necessary, the Jam required familiarization with emerging technologies which were described in online materials made available to participants prior to the event.

Individual contributions to the Jam came in the form of "*postings*," or messages in reply to other contributors and to questions posed under a moderated topic area. As shown in Figure 1, group

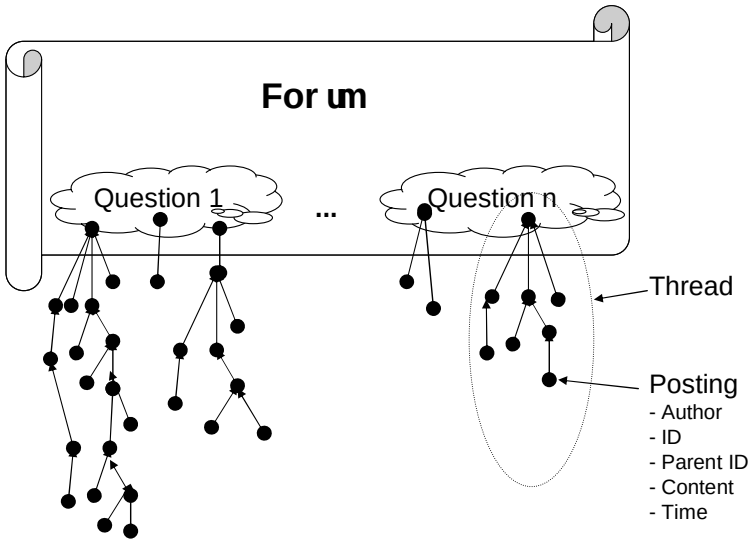

Figure 1: Relationship between postings, threads, questions and forums in both Jam phases.

For five weeks following Phase 1 of the Innovation Jam, a multi-discipline, international cross-IBM team analyzed more than 37,000 Phase 1 posts to identify the most promising suggestions, resulting in 31 identified topics or "*big ideas*" as listed in Table 4. Taking the

raw ideas from Phase 1 and transforming them into real products, solutions and partnerships to benefit business and society was the focus of Innovation Jam Phase 2, September 12-14, 2006, which involved more focused sessions where participants refined ideas.

In light of the discussion in Jam Phase 2, internal teams drawing on a broad range of subject-matter expertise continued to discuss and shape the key emerging ideas, evaluating them in terms of their technological innovation as well as their potential impact on society and business. Based on these discussions, a final list of ten leading ideas or "*finalists*" were identified to receive funding for development over the next two years. The ten finalists were:

1. **3-D Internet**: Establish the 3-D Internet as a seamless, standards-based, enterprise-ready environment for global commerce and business.

2. **Big Green innovations**: Enter new markets by applying IBM expertise to emerging environmental challenges and opportunities.

3. **Branchless Banking**: Profitably provide basic financial services to populations that don't currently have access to banking.

4. **Digital Me**: Provide a secure and user-friendly way to seamlessly manage all aspects of my digital life - photos, videos, music, documents, health records, financial data, etc. - from any digital device.

5. **Electronic Health Record System**: Create a standards-based infrastructure to support automatic updating of - and pervasive access to healthcare records.

6. **Smart Healthcare Payment System**: Transform payment and management systems in healthcare system

7. **Integrated Mass Transit Information System**: Pursue new methods to ease congestion and facilitate better flow of people, vehicles and goods within major metropolitan areas.

8. **Intelligent Utility Network**: Increase the reliability and manageability of the world's power grids.

9. **Real-Time Translation Services**: Enable innovative business designs for global integration by removing barriers to effective communication, collaboration and expansion of commerce.

10. **Simplified Business Engines**: Deliver the "Tunes" of business applications.

Table 1: Summary statistics for the two phases conducted in Innovation Jam

Summary Statistics	**Phase 1**	**Phase 2**
No. of Messages	37037	8661
No. of Contributors	13366	3640
No. of Threads	8674	254
No. of Threads with no response	5689	0
No. of Threads with $\leq$10 responses	2673	60
No. of Threads with $\geq$100 responses	56	12

While recognizing that significant human processing took place in the course of evaluating Jam data, our goal is to see if we can identify factors that would have been predictive of the Jam finalists, perhaps suggesting ways to help make processes for future Jams less manually intensive.

3.2 Overview of the Jam Characteristics

As mentioned above, the Innovation Jam was conducted in two phases that were separated by a period of less than 2 months. Table 1 summarizes some of the basic statistics of these

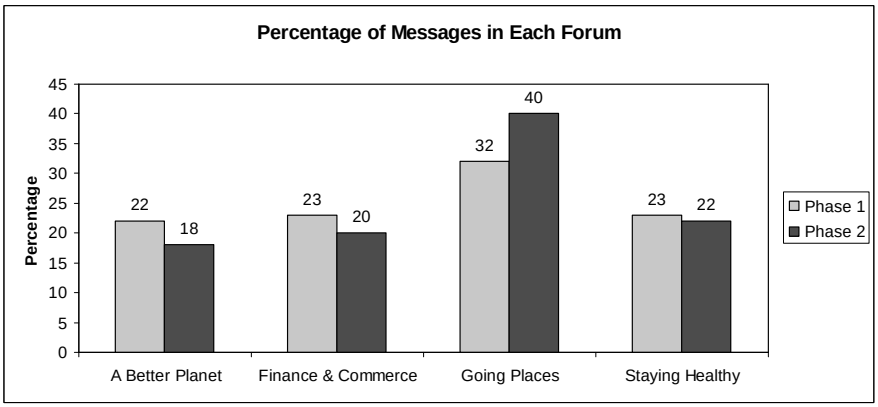

Figure 2: Percentage of messages in each forum for Phase 1 and Phase 2.

two phases. In both phases, all the threads belonged to one of the following four forums: (1) Going Places, (2) Staying Healthy, (3) A Better Planet and (4) Finance and Commerce

Figure 2 gives the percentage of messages in each of the above mentioned forums during Phase 1 and Phase 2. We can see that topics related to "Going Places" received relatively more attention during Phase 2. Percentage of contributors who responded more than 1-20 times during both phases is shown in Fig. 3. Considering the fact that the numbers of contributors are 13366 in Phase 1 and 3640 in Phase 2, it is interesting to note that these percentages are very similar for both phases. For example, percentage of contributors who responded at least 4 times is 18% for Phase 1 and 16% for Phase 2.

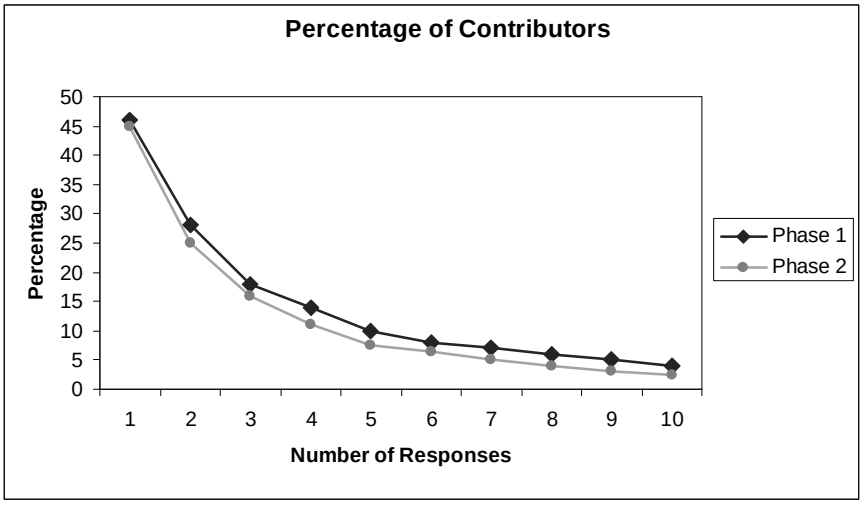

Figure 3: Percentage of contributors who responded more than 1-20 times during Phase 1 and Phase 2.

3.3 Sources of Data

In the case of the IBM Innovation Jam, we have access to unique and highly-diverse sources of high quality data to be used in addressing our learning challenge. We now briefly review the data sources and types we have available. In the succeeding sections we will describe the data itself and our analytical approaches.

These data sources are:

1. **The text of the threads itself.** From analyzing the text we can find similarity between threads, understand how tight the discussion in each thread was, identify the keywords differentiating between threads.

2. **The social network structure of threads and the whole Jam.** Within each thread, we can analyze the structure of the discussion, and collect statistics such as

how many "leaves" (postings with no response) there were, how deep is the typical discussion in the thread, etc. Since we have unique identifiers for all contributors, we can also analyze the connection between threads through common contributors, the variety of contributors in each thread (e.g, messages per poster).

3. **The organizational relationships between the contributors.** Since the vast majority of contributors were IBM employees, we can make use of the online directory of worldwide IBM employees (known as Blue Pages), to capture the geographic and hierarchical relationships between the contributors as well as in each thread, in each *Big Idea*, etc.

Since a prevalent hypothesis is that a major advantage of the Jam is that it brings together people from different parts of IBM, and different geographical locations, who would otherwise be unlikely to interact; and that such interactions between diverse groups are likely to lead to new insights and innovations, data in this last category may be of particular interest in our analysis.

4 Social Network and Dynamics in the Jam Interactions

Let us take a closer look at the social aspect of the Jam domain and in particular how the network of interactions between contributors evolves over time. Figure 4 shows the number of postings per hour over the 3 days of the Phase 1. The plot shows clear seasonality of a 24-hour cycle, after an initial spike within the first two hours. The hourly count of contributions remains fairly stable over the 3 day period. In the sequel we will consider the social network of contributors where every node is a contributor and a directed link

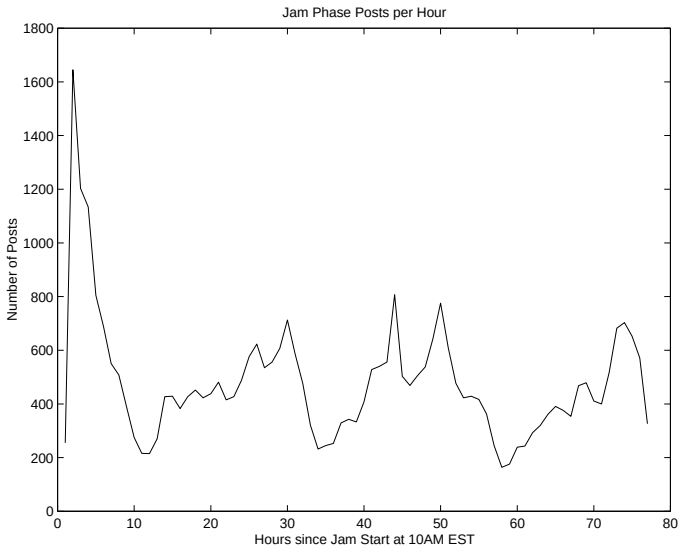

Figure 4: Number of postings over time during Jam Phase 1.

from person A to person B is present if A directly responded to B. We can extract this relationship from the posting identifiers and the provided identifier of the parent posting. Here, social behavior is modeled as a network in an approach similar to [1].

The resulting social Jam network is shown for a number of points in time (2 hours, 3 hours, 4 hours and 10 hours after start of Jam) in Figure 5. The graph layouts are generated using Pajek [4].

Figure 6 shows a histogram of the organizational distances between IBM contributors who posted replies to other IBM contributors. (Posts connecting IBM contributors and external people were not considered.) Here, organizational distance between two contributors is determined by traversing the reporting structure until a common person is found, and counting the number of nodes (people) encountered in the complete traversal [1]. The average distance of 11.6, and the wide distribution of distances, suggests that the Innovation Jam

[1]The IBM reporting structures are a forest of trees, with each tree representing the reporting structure in one geographic location, e.g. US or China. If two employees are from different countries, they are disconnected in the reporting structures. In this case, we incremented this distance by 2 to bridge geographic disconnection.

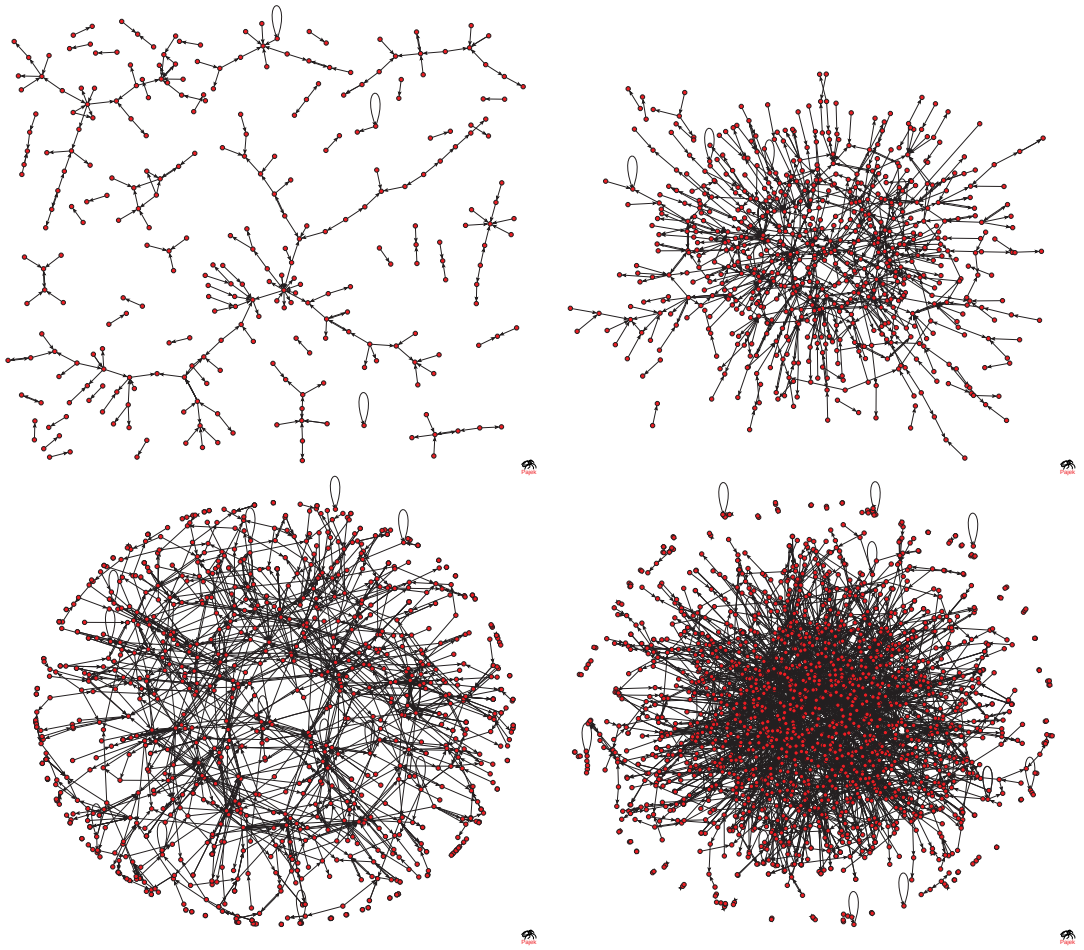

Figure 5: Evolvement of the social Jam network for 2, 3, 4 and 10 hours after the Phase 1 start. The initial network has after 2 hours still a number of independent components that most likely reflect the thread structure of the Jam. However, already after 3 hours the Jam population is fairly interconnected and only a few very small independent components remain. This trend continues until after 10 hours the network structure compacts into a tight ball with a number of small peripheral components. Given the linear trend in the population and the rather constant rate of postings, the overall density (number of present links over number of possible links) of the social network is exponentially decreasing.

was successful as a forum for creating interactions between people who otherwise might have been unlikely to interact.

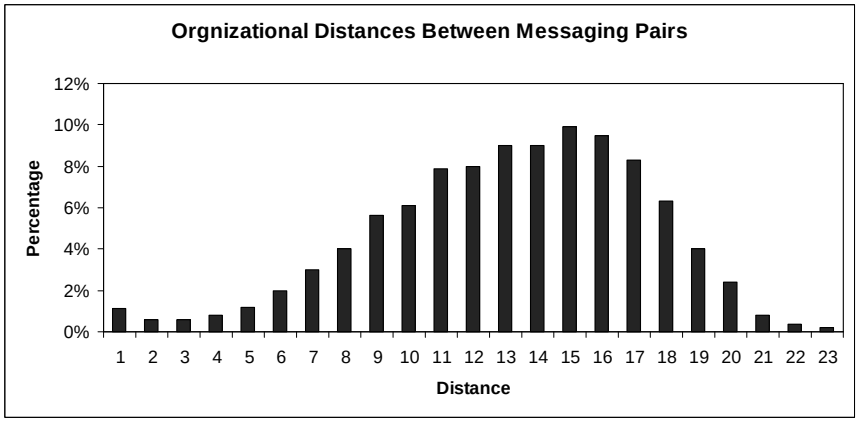

Figure 6: Organizational distances between IBM contributors who posted replies to other IBM contributors

4.1 Browsing or Focused contribution?

The observed network structure suggests that the individual Jam contributors are not focused on a single thread but rather seem to 'browse' between threads and topics. If individual contributors were contributing to a single thread we would expect the network to show a number of loosely connected islands (corresponding to threads) with high intra-connectivity. As a first consideration we estimate the probability of a repeating contributor to post to a new thread, where new is defined as a thread to which he has never posted. This probability is calculated for each contributor as the number of contributed posts minus one (the first posting is by definition to a thread that the contributor has not posted in before) divided by the number of different threads in which he or she posted.

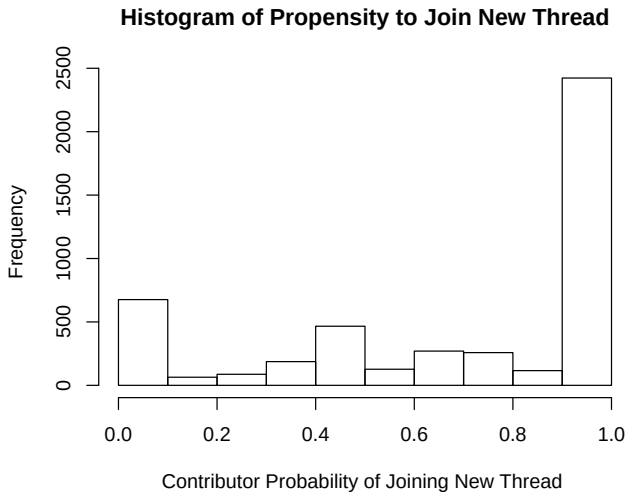

Figure 7: Histogram of the propensity of contributors to post to a thread they have not posted to before rather than positing to a thread they did contribute to in the past.

And indeed, the average probability is surprisingly high at 62%. The histogram in Figure 7 shows that a large number of Jam contributors ventured into multiple threads. The large spike around probability 1 is caused by contributors with only 2 postings in two different threads. However, the scatter plot in Figure 8 reveals that there is no clear functional dependence between the number of postings of a contributor and his probability of contributing to multiple threads.

4.2 Network Correlation

One of the questions that remains open is whether the social network changes structurally over the time of the Jam (and is thereby more or less accidental) or reflects some stable social bonds. To address this question we will measure the similarity of the social network over time. In particular, we will compare the networks from adjacent time windows with a certain number of posts. We define the similarity coefficient SC similar to the Jaccard coefficient [12] of two networks with identical nodes N but varying edgesets E_1 and E_2

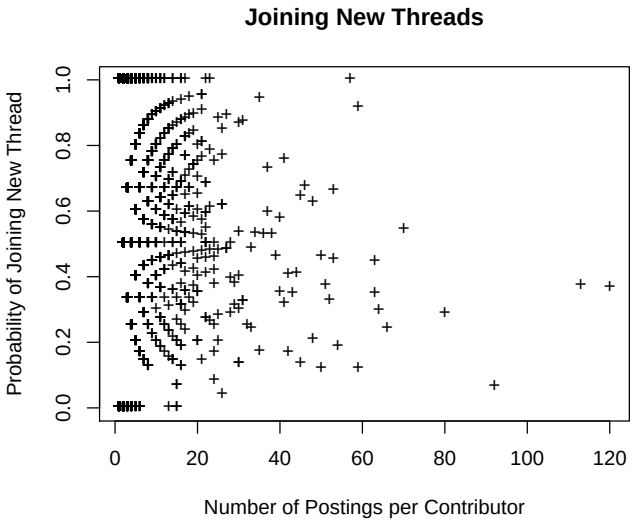

Figure 8: Histogram and scatterplot of the propensity of contributors to post to a thread they have not posted to before rather than positing to a thread they did contribute to in the past.

simply as the number of identical edges over the number of total edges:

$$SC(E_1, E_2) = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2} \tag{1}$$

For a window size of n postings, we calculate the average similarity coefficient ASC_n over all pairs of non-overlapping sequential networks derived from n consecutive postings. In order to make this analysis relevant we need to determine a meaningful baseline. In Figure 9 we show the ratio of ASC_n over a random $ASC*_n$ that is calculated from pairs of random networks. We construct two random networks with identical general properties by randomly (without replacement) picking $2*n$ edges from the total social Jam network in Phase 1. We observe from the graph that the similarity between actual networks of sequential windows is always higher than the similarity of random network pairs. However, the relative similarity changes as a function of the window size. For very small windows of only 50 postings, the actual graphs have much less similarity than for a window size of 200 at which the

Figure 9: Ratio of the autocorrelation of social networks from Phase 1 of a given size to the expected correlation of two random networks.

graph seems to peak. For windows larger then 200 the graphs steadily declines. This phenomenon can be explained in terms of actual time, 50 postings correspond to about 20 seconds. Writing a meaningful posting will typically take more time than this. So for very short periods the similarity is low because there is not enough time to write another post. The maximum similarity is around 1 minute, giving the participant plenty of time to post another reply. Beyond a time frame of 5 minutes however, the interest in a particular topic seems to decrease and the participant will post elsewhere or leave altogether.

4.3 Social Dynamics of Innovation Jam

When analyzing an event like the Innovation Jam, it is often useful to explore the social structure of the participants thereof. A key question to ask in the context of such a brainstorm is whether or not groups or communities formed in a short time frame — specifically, did people post ideas without elaborating and discussing together, or was the three-day period long enough for communities to form and collaborate on ideas?

The social network itself can play a significant role in clarifying communication patterns between participants. A useful approach is through a dyad census [24], which counts the types of edges between nodes. In such a case, all edges are isomorphic to three groups: a non-existent edge (node A and node B have no connection), a directed edge or arc (node A replied to node B but not vice versa), and a mutual edges (node A replied to node B, and node B replied to node A). In the case of a threaded discussion, the ratio of mutual ties to individual arcs can be useful in showing whether pairs of individuals replied to each other throughout the process.

The actual density of the social network throughout the entire three-day Jam period is 0.00014. Such a low density results in a large number of non-existent edges, so rather than analyzing the results of a dyad census directly, the ratio of mutual edges to non-mutual ones was computed over two types of networks. Both networks were built by separating the Jam data into 12 periods, representing approximately 6 hours of threaded discourse. The first network was analyzed cumulatively, by aggregating the various posts up to specific periods, while the other only had individual periods analyzed without incorporating earlier information.

When compared to a random network where the probability of an edge between two nodes is equal to the density D, the likelihood of a reciprocal tie is D^2. As such, the expected number of mutual edges at the end of the entire three-day period is extremely low, at about 2. However, the observed network spanning the entire three-day period has many more mutual edges — approximately 8.4% of all dyads are mutual, which represents over 1800 edges. The variation within the 6-hour intervals shows that such relationships required larger time frames to actually develop, rather than taking place within specific 6-hour time

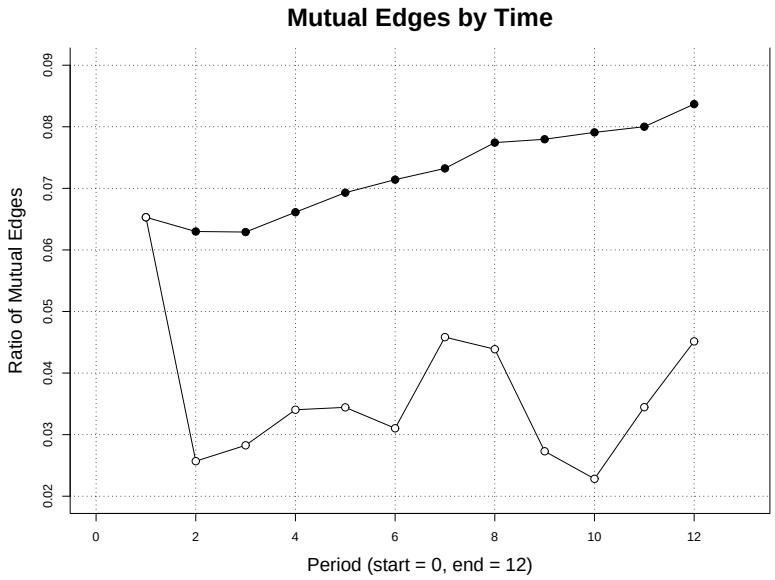

Figure 10: Ratio of mutual edges to all edges within the Innovation Jam. White points represent 6-hour intervals, while black points represent the cumulative state of the network from the start of the Jam until the specific point in time.

frames.

However, community formation within social networks must involve more than two individuals replying to each other. To measure group formation within a network, it is useful to measure the clustering coefficient of a network. At the node level, the clustering coefficient represents the probability that the neighbors of a node n have ties between each other. The definition of the clustering coefficient [25] for a node n in a directed network is

$$C(n) = \frac{E(n)}{k(n)(k(n)-1)} \tag{2}$$

where $E(n)$ is the number of edges among the neighbors of n, and $k(n)$ is the number of neighbors of n. Note that in a directed network, neighbors include nodes that are connected to n through both in- and out-degrees. Thus, a clustering coefficient value of 0.25 for a node n states that 25% of the possible edges between neighbors of n actually exist. The

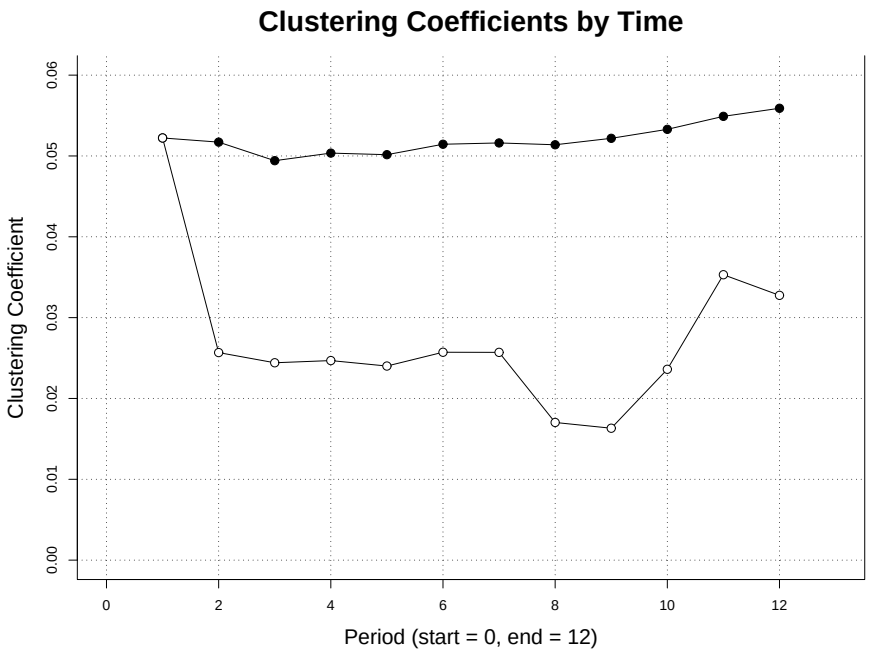

Figure 11: Clustering coefficient of the social network of Jam participants. White points represent 6-hour intervals, while black points represent the cumulative state of the network from the start of the Jam until the specific point in time.

clustering coefficient can be extended to an entire network by simply taking the average of the individual nodes. It should be noted that when doing so, nodes with 0 or 1 degree are ignored.

The cumulative networks clustering coefficient remains fairly stable throughout the entire three-day Jam, showing that the amount of clustering of individuals did not change while new people involved themselves in discussions and others continued to participate. Throughout the 6-hour intervals, the clustering coefficient drops off in the shorter periods, again showing that that edges beyond the 6-hour intervals seem to complete the network and maintain the community-like structure of the network. This could be caused by when participants actually post and reply to threads — depending on the time zone or personal preferences, they may be prefer to post at a specific time of day or outside of certain 6-hour intervals.

The clustering coefficient itself hovers around 0.052 for the cumulative network, implying a high level of clustering compared to what would be expected from a random network, where simulations show an expected coefficient of 0.00016 (standard deviation of 0.00005). In comparison, Ebel et al. [8] observed a clustering coefficient of 0.0344 in e-mail networks. Strogatz [21] observed coefficients of 0.088, 0.199, and 0.588 for networks of biomedical researchers, movie actors, and company directors, respectively.

Based on the clustering coefficients and ratios of mutual edges, it appears the network exhibits some properties of community formation and discussions. As such, one can see that rather than simply acting as a bulletin board-like tool where individuals could post ideas and leave, clustering and discussions can be observed. It is important to note, however, that about 70% of participants only participated by posting once or twice rather than discussing topics regularly, though the results above show that this did not deter others from in-depth brainstorming. Furthermore, it is possible to gain more insight about the social structure of the Jam itself by analyzing the actual discussions themselves.

5 Unsupervised Content Analysis

The high-level challenge of our analysis of the Innovation Jam data is to identify the keys to success of such an endeavor, in particular, what are the characteristics of discussion threads that lead to innovative and promising ideas? The major differences between the Jam data and a typical forum are: a) the topics are more concentrated and controlled; b) the contributors are mostly from one organization, and therefore share similar concepts on basic values and what are the "great" ideas; c) the discussion time spans a shorter time.

As in every learning problem, there are two general approaches that can be taken to address this challenge:

- **The supervised learning approach.** If we could go back and label discussion threads as *successful* or *unsuccessful*, we could then investigate and characterize the features differentiating between the two classes, and hypothesize that these are the features that lead to success. As we discuss below, we have utilized the selection of big ideas from the Jam as finalists for funding for labeling, and attempted to correlate the various features with this selection, with limited success so far.

- **The unsupervised learning approach.** The idea here is to concentrate our effort on characterizing and categorizing the discussion threads in terms of their typical profiles, or groups of distinct typical profiles. While this analysis may not lead directly to conclusions on which profiles represent *successful* Jam threads, it can be an important step towards hypothesis generation about success. Furthermore, it can be used as an input to discussions with experts and to design of experiments to test the success of the different thread types in generating innovation. We describe the promising results of unsupervised learning on Jam text features below.

We discuss the unsupervised approach in this section, and defer the discussion of supervised techniques to Section 6.

5.1 Data Preprocessing

To analyze the content of the jam postings, we preprocess the text data and convert them into vectors using bag-of-words representation. More specifically, we put all the postings

within one thread together and treat them as one big document. To keep the data clean, we remove all the threads with less than two postings, which results in 1095 threads in Phase 1 and 244 threads in Phase 2. Next, we remove stop words, do stemming, and apply the frequency-based feature selection, i.e. removing the most frequent words and those appearing less than 2 times in the whole collection. These processes results in a vocabulary of 10945. Then we convert the thread-level documents into the feature vectors using the "ltc" TF-IDF term weighting [5].

5.2 Clustering algorithm

Our objective of the unsupervised analysis is to find out what are the overlapping topics in Phase 1 and Phase 2, i.e. the topics that discussed in Phase 1 have been picked up by Phase 2, which can be seen as a potential indicator of "finalists" of ideas for funding. Therefore when we cluster the threads from Phase 1 and Phase 2, an optimal case is that we can find three types of clusters: (1) the clusters that mostly consist of threads in Phase 1 (2) those mostly composed of threads in Phase 2; and (3) the clusters with the threads in both phases, which help us examine if they are the potential finalists for funding. Several clustering algorithms have been investigated, including K-means, hierarchical clustering, bisecting K-means and so on [13]. The results from different clustering algorithms are similar and therefore we only discuss the ones using the complete-linkage agglomerate clustering algorithm. For implementation, we use the open source software CLUTO [2].

[2]http://glaros.dtc.umn.edu/gkhome/views/cluto

5.3 Clustering Results

As discussed above, we use the document clustering algorithms to analyze the threads in Phase 1 and Phase 2. In the experiment, we preset the number of clusters to 100. Several interesting observations can be made by examining the clustering results: (1) Phase 1 to Phase 2: since we are interested in finding out the overlapping topics between the threads in Phase 1 and those in Phase 2, we plot the histogram on the number of threads from Phase 2 in each cluster in Figure 12. From the results, we can see that the majority of the clusters (around 70%) only contain the threads in Phase 1, which indicate that the topics in phase 2 are only a subset of those in Phase 1 and there are no new topics in Phase 2. This agrees well with the process of the Jam, i.e. a subset of the topics discussed in Phase 1 are selected and used as discussion seed in Phase 2. (2) Phase 2 to Finalist ideas: we further examine the topic overlapped between the threads in Phase 2 and those selected as successful finalist ideas by going through the clusters with the most threads from Phase 2. From Table 2, we can see an impressively direct mapping from the top-ranked clusters (by the number of threads from Phase 2) to the finalist. For example, the cluster with the largest number of threads from Phase 2 is shown in the first line. It seems to concentrate on the topics about "patients", "doctors" and "healthcare", which agrees well the main theme in one of the finalist ideas, i.e. "Electronic Health Record System". Another example is the cluster devoted to the idea of "Digital Me". Its descriptive words are "dvd", "music", "photo" and so on, which clearly reflects the theme about providing a secure and user-friendly way to seamlessly manage photos, videos, music and so on.

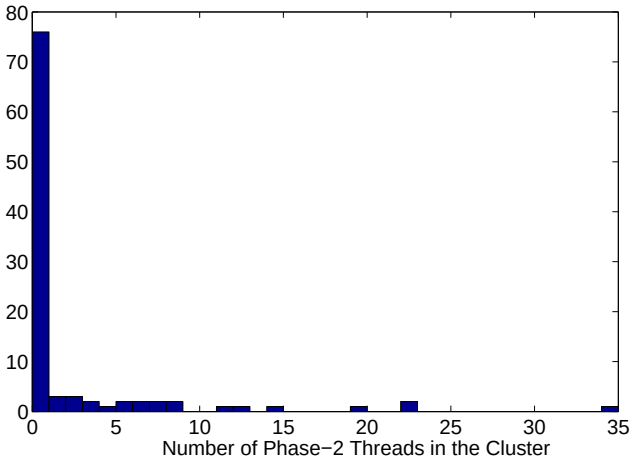

Figure 12: Histogram of the number of Phase 2 threads in the 100 clusters

5.4 Topic Tracking

In this section, we explore the evolution of discussion concentration over time. In particular, we are interested in answering the questions: What are the possible characteristics of a focused discussion, and are there any patterns as to when and how the focused discussion can happen? The three largest threads were selected for analysis, with the first discussing "Sustainable energy systems", the other on "Digital entertainment" and the third on "Global travel" with 622, 405 and 776 posts respectively.

To determine whether a series of posts have a focused discussion or not, we compare the content similarity as well as the number of unique contributors. We preprocess the texts in each post using the bag-of-words representation described before, and then calculate the averaged cosine similarity between 10 adjacent posts in a time-based consecutive order. In addition, we calculate the statistics of the number of unique contributors in a way that penalizes discussions involving very few contributors (to avoid random chatting between two persons). Figure 13 shows the discussion concentration with content similarity and the number of unique contributors. There seems to be a general trend common in all the

three threads, that is, at the beginning of the Jam there are many posts on multiple topics, then these topics evolve as time goes, and finally one or more of them lead to a focused discussion. Usually these effective discussions appear at around 20 to 30 hours after the Jam starts.

To further examine the effectiveness of our concentration measurement, we show an example of 5 posts with the highest averaged cosine-similarity score and the most unique contributors in the "digital entertainment" thread. These posts occur at around 27 hours after the jam starts as identified in Figure 13).

1. "... Going to the movies is a social experience. It is about enjoying the whole experience: large screen, popcorn, people ..."
2. "... if you want to experience that you might want to go to disneyland to see/feel 'honey I shrunk the audience"'
3. "The possible future development in entertainment will be the digital eye glasses with embedded intelligence in form of digital eye-glasses. The advantages for users would be: the screen size and the 'feeling' to be inside the action ..."
4. " ... Really big screens accompanied by great sound and special lighting effects would really enhance the experience and make it different from renting a movie at home ..."
5. "It would be nice if multiple movies could play on the same screen and the audience would wear special glasses so they could see only the movie they payed for. This would reduce the need for multiple theater auditoriums in a single location. "

We can see that the discussion is significantly focused on multiple ways of improving the theater experience.

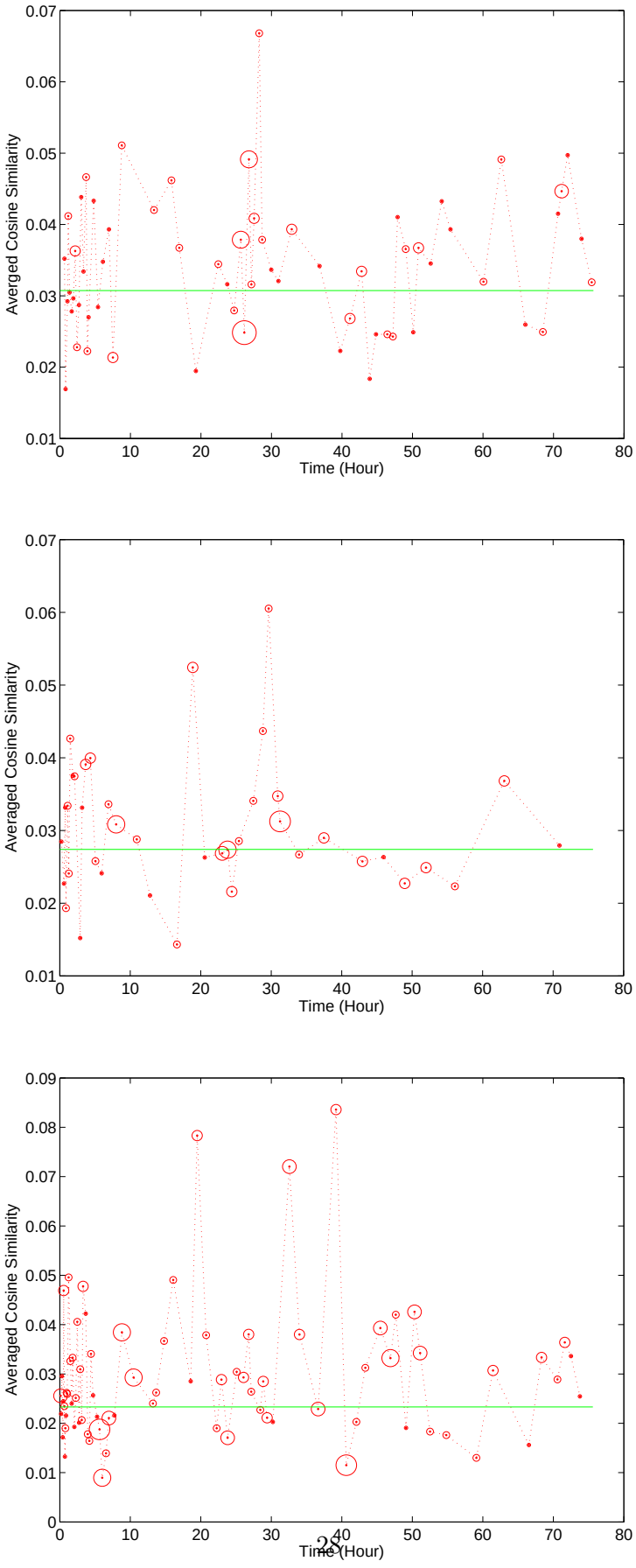

Figure 13: Plot of averaged similarity between 10 posts in most common threads; the size

6 Finding Great Ideas: Supervised Analysis

Our supervised learning approach concentrated on trying to identify what makes a "big idea" into a "finalist" for funding. We therefore calculated 18 features from the various jam data sources, describing the interactions and discussions in phase 2 of the Jam. These features can be classified according to the three categories of data from which they were derived:

1. **Topological Features**: Features T1-T8 described in Table 3 correspond to topological features. These features will give some basic intuition about the Phase 2 of the Innovation Jam. It contains information regarding the topology of the messaging including number of messages, number of contributors, number of questions in a given idea, number of responses for each question and so on. Column T8 corresponds to the interconnection of contributors between these ideas. It gives the number of times that the contributors of a given idea participated in other ideas. The contributors are weighted based on their contribution in the given idea.

2. **Contextual Features**: Features C1-C5 described in Table 3 correspond to contextual features. These features are computed based on the bag-of-words representation of all the messages belonging to a single thread. The pairwise cosine similarity measure is computed between all possible pairs of threads with more than one message in a particular big idea. Some basic statistics like the mean, standard deviation, maximum and minimum of these scores are considered as features.

3. **Organizational Features**: Features O1-O5 described in Table 3 correspond to organizational features. Basically, organizational distance between two contributors can

be computed by traversing a 'management tree' where each node corresponds to a person and its parent node corresponds to the person to whom he reports to. The distance between two contributors can be obtained by *climbing up* each of the trees until a common person is found[3]. Sometimes, two contributors might not have any common personnel in the reporting structure. In those cases, both the lengths of the reporting structure for the two contributors are added and the total is incremented by 2 (considering the fact that people in the topmost position in the ladder are somehow connected by another imaginary layer). Again, some basic statistics are computed as described above.

The values of these eighteen features are computed for all the 31 big ideas (Table 4). We also associate a *label* field with each big idea, indicating whether or not it was chosen as a "finalist" for funding. Hence, we can treat this as a supervised learning problem and we can use the labeling to identify the most informative features.

Testing the features for association with selection for funding

We investigated the correlation between our 18 features and the response variable — whether or not each "big idea" was selected as a finalist for funding. We applied a parametric t-test, and two non-parametric tests (Kolmogorov-Smirnov and Mann-Whitney, [7]) to test the hypothesis of a difference in the distributions P(feature|selected) and P(feature|not selected) for each of the 18 features. The results (Table 3) demonstrate that there is no evidence that any of the features carries significant information about the selection process.

[3]For few contributors, it was difficult to obtain the organizational hierarchy information. These cases were eliminated during the computation.

[3]Excluding the questions with less than 10 responses

[4]Threads containing more than one message

[5]For only those contributors whose organizational information was available

Finalist Ideas for Funding	**P1**	**P2**	**Descriptive Stemmed Words**
Electronic Health Record System	49	35	patient, doctor, healthcar, diagnosi, hospit, medic, prescript, medicin, treatment, drug, pharmaci, nurs, physician, clinic, blood, prescrib, phr, diagnost, diseas, health
Digital Me	26	23	scrapbook, music, dvd, song, karaok, checker, entertain, movi, album, content, artist, photo, video, media, tivo, piraci, theater, audio, cinema
Simplified Business Engines	26	23	smb, isv, back-offic, eclips, sap, mashup, business-in-a-box, invoic, erp, mgt, oracl, app, salesforc, saa, host, procur, payrol, mash, crm
Integrated Mass Transit Information System	59	20	bus, congest, passeng, traffic, railwai, commut, rout, lane, destin, transit, journei, rail, road, vehicl, rider, highwai, gp, driver, transport
Big Green innovations	27	13	desalin, water, rainwat, river, lawn, irrig, rain, filtrat, purifi, potabl, osmosi, contamin, purif, drink, nanotub, salt, pipe, rainfal, agricultur
3-D Internet	22	12	password, biometr, debit, authent, fingerprint, wallet, finger, pin, card, transact, atm, merchant, reader, cellular, googlepag, wysiwsm, byte, userid, encrypt
Intelligent Utility Network	23	9	iun, applianc, peak, thermostat, quickbook, grid, outag, iug, shut, holist, hvac, meter, heater, household, heat, resours, kwh, watt, electr, fridg
Branchless Banking	11	9	branchless, banker, ipo, bank, cr, branch, deposit, clinet, cv, atm, loan, lender, moeni, withdraw, teller, mobileatm, transact, wei, currenc, grameen
Real-Time Translation Services	33	5	mastor, speech-to-speech, speech, languag, english, nativ, babelfish, translat, troop, multi-lingu, doctor-pati, cn, lanaguag, inno, speak, arab, chines, barrier, multilingu

Table 2: The mapping from the clusters with the most threads in Phase 2 to the finalist ideas. P1 and P2 are the number of threads in the cluster from Phase 1 and from Phase 2 respectively.

The last feature, *Minimum pairwise distance between the contributors*, results in a p-value that is smaller than 0.05 for a couple of tests, but given the amount of multiple comparisons we are doing, this can by no means be taken as evidence of real association. Thus we can conclude that our 18 features fail to capture the "essence" of the Jam as it pertains to the finalist funding decisions. Discovering and formalizing this essence remains a topic for future work.

7 Conclusion

Our broad objective in this work is to apply social network analysis and machine-learning-based techniques to data obtained from moderated, online forums like Innovation Jam with the immediate goal of identifying aspects of these discussions that lead to innovative ideas. This is a particularly challenging task. We have applied both supervised and unsupervised approaches to the Jam data, which includes labeling of the most innovative ideas based on human-expert insight. Examination of a range of features drawn from analysis of the topology of the discussion, the context of the discussion, and the organizational diversity of the participants did not yield strong statistical evidence concerning how innovative ideas evolve.

Although this supervised study was not as successful as hoped, this work has shown that the Jam data does exhibit some ordering both in terms of the social structure and discussions themselves. It has been observed that within short time frames (i.e. within minutes), discussions between individuals can be observed, and that over time, (i.e. hours and days) general discussions tend to become more focused and specific.

The Innovation Jam was a unique implementation of a threaded discussion due to its corporate focus, short time frame, and use of moderators. It is encouraging to see that even in such a short period, collaboration can be observed and people can begin working together to generate novel ideas. Much work is left in extending our use of the different data types in both supervised and unsupervised learning, and in identifying the key characteristics — or combination of characteristics — that lead to success.

Acknowledgments

We would like to thank Cathy Lasser, Samer Takriti, Jade Nguyen Strattner for general discussions and insights about the IBM Innovation; Noel Burke, Selena Thomas for help finding information on the IBM Innovation Jam process; William Tuskie and Scott Sprangler for providing the Jam data; Cathy Lasser, David Yaun for their critical review and discussion of our conclusions; and John Thomas and Kate Ehrlich for insights regarding social networks, and socio-technical systems.

References

[1] R. Agrawal, S. Rajagopalan, R. Srikant, and Yirong Xu. Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th International Conference on World Wide Web*, pages 529–535, 2003.

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[3] Micheal Ashworth and Kathleen Carley. Who you know vs. what you know: The impact of social position and knowledge on team performance. *Journal of Mathematical Sociology*, 30, 2006.

[4] V. Batagelj and A. Mrvar. Pajek - program for large network analysis. *Connections*, 2(21):47–57, 1998.

[5] C. Buckley, G. Salton, and J. Allan. The effect of adding relevance information in a relevance feedback environment. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. Springer-Verlag, 1994.

[6] Ronald S. Burt. Structural holes and good ideas. *American Journal of Sociology*, 110(2):349–399, 2004.

[7] W.J. Conover. *Practical nonparametric statistics*. New York: John Wiley and Sons, 1971.

[8] H. Ebel, L.I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Physical Review*, 66(3):35–103, 2002.

[9] S.A. Golder and J. Donath. Social roles in electronic communities. *Internet Research*, 5:19–22, 2004.

[10] C. Halverson, J. Newswanger, T. Erickson, T. Wolf, W. A. Kellogg, M. Laff, and P. Malkin. World jam: Supporting talk among 50,000+. In *Proceedings of the European Conference on Computer-Supported Cooperative Work (ECSCW)*, 2001.

[11] J. Hempel. Big blue brainstorm. *Business Week*, 32, August 2006.

[12] P. Jaccard. Eťude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Societ́eV́audoise des Sciences Naturelles*, 37:547–579, 1901.

[13] A.K. Jain and R.C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[14] Henry C. Lucas Jr., Swanson E. Burton, and Robert Zmud. Implementation, innovation, and related themes over the years in information systems. *Journal of the Association for Information Systems*, 8(4), 2008.

[15] Jon Kleinberg. The Small-World Phenomenon: An Algorithmic Perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.

[16] P. Kolari, T. Finin, K. Lyons, Y. Yeshaa, Y. Yesha, S. Perelgut, and J. Hawkins. On the structure, properties and utility of internal corporate blogs. In *International Conference on Weblogs and Social Media*, 2007.

[17] C. Lasser. Discovering innovation. In *IEEE International Conference on e-Business Engineering (ICEBE)*, 2006.

[18] Jay Liebowitz. *Social Networking: The Essence of Innovation*. The Scarecrow Press, Inc., 2007.

[19] Wendy Olphert and Leela Damodaran. Citizen participation and engagement in the design of e-government services: The missing link in effective ict design and delivery. *Journal of the Association for Information Systems*, 8(9), 2008.

[20] Manoj Parameswaran and Andrew B. Whinston. Research issues in social computing. *Journal of the Association for Information Systems*, 8(6), 2008.

[21] S.H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.

[22] J. Travers and Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.

[23] T.C. Turner, M.A. Smith, D. Fisher, and H.T. Welser. Picturing Usenet: Mapping Computer-Mediated Collective Action. *Journal of Computer-Mediated Communication*, 10(4), 2005.

[24] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge: Cambridge University Press, 1994.

[25] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):409–10, 1998.

Index	Description of the Feature	t-test	K-S test	M-W test
T1	Total Number of messages for a particular big idea.	0.58	0.99	0.67
T2	Total Number of messages which didn't receive any further response.	0.61	0.97	0.60
T3	Total Number of contributors.	0.92	0.94	0.95
T4	Forum Number.	0.86	1	0.90
T5	Total Number of questions asked in that particular idea.	0.70	0.91	0.71
T6	Mean of the number of messages for all questions [4].	0.96	0.69	0.82
T7	Standard deviation of the number of messages for all questions [4] .	0.53	0.90	0.66
T8	Weighted number of overlapping contributors involved in other big ideas.	0.91	0.88	1
C1	Mean of the pairwise cosine similarity scores between the threads [5].	0.31	0.70	0.34
C2	Standard deviation of the pairwise scores between the threads [5].	0.40	0.29	0.28
C3	Total number of pairwise scores between all threads.	0.52	0.85	0.46
C4	Maximum pairwise score between the threads.	0.38	0.91	0.90
C5	Minimum pairwise score between the threads.	0.94	0.84	0.79
O1	Average pairwise distance between the contributors within a big idea [6].	0.62	0.66	0.54
O2	Standard deviation of the pairwise distances between the contributors [6].	0.91	0.94	0.97
O3	Total number of pairwise distances between all the contributors involved.	0.93	0.90	0.98
O4	Maximum pairwise distance between the contributors.	0.64	0.85	0.59
O5	Minimum pairwise distance between the contributors.	0.046	0.29	0.046

Table 3: Description of 18 different features used in the analysis of Innovation Jam.

Table 4: Summary of 31 big idea names obtained from the analysis of Phase 2 and label indicating whether they were under the finalists selected for funding.

Big Idea	**Funded**
Rail Travel for the 21st Century	0
Managed Personal Content Storage	1
Advanced Safecars	0
Health Record Banks	1
The Truly Mobile Office	0
Remote Healthlink	0
Real-Time Emergency Translation	1
Practical Solar Power Systems	0
Big Green Services	1
Cellular Wallets	0
Biometric Intelligent Passport	0
Small Business Building Blocks	0
Advance Traffic Insight	0
3-D Internet	1
Branchless Banking for the Masses	1
e-Ceipts	0
Digital Entertainment Supply Chains	0
Smart Hospitals	0
Business-in-a-box	1
Retail Healthcare Solutions	0
Digital Memory Saver	0
Intelligent Utility Grids	1
Cool Blue Data Centers	0
Water Filtration Using Carbon Nanotubes	0
Predictive Water Management	0
Sustainable Healthcare in Emerging Economies	0
Bite-Sized Services For Globalizing SMBs	0
Integrated Mass Transit Information Service	1
Smart-eyes, Smart-insights	0
Smart Healthcare Payment Systems	1
Advanced Energy Modelling and Discovery	0