RC24277 (W0706-026) June 5, 2007
Computer Science

IBM Research Report

Looking for Great Ideas: Analyzing the Innovation Jam

Mary Helander, Rick Lawrence, Yan Liu, Claudia Perlich,
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 - Haifa - India - T. J. Watson - Tokyo - Zurich

Looking for Great Ideas: Analyzing the Innovation Jam

Mary Helander, Rick Lawrence, Yan Liu,
Claudia Perlich, Chandan Reddy, Saharon Rosset
IBM T.J. Watson Research Center
P. O. Box 218
Yorktown Heights, NY 10598
{helandm, ricklawr, liuya, perlich, crkarrem, srosset}@us.ibm.com

ABSTRACT

We discuss the Innovation Jam that IBM carried out in 2006, with the objective of identifying innovative and promising "Big Ideas" through a moderated on-line discussion between IBM worldwide employees and external contributors. We describe the data available and investigate several analytical approaches to address the challenge of understanding "how innovation happens" and to facilitate the success of future Jams. We demonstrate the social network structure of data and its time dependence, and discuss the results of both supervised and unsupervised learning applied to this data.

1. INTRODUCTION

Social network analysis is the mapping and measuring of relationships and flows between people, groups, organizations or other information/knowledge processing entities. There have been tremendous work on the social network study over the past century [9, 1, 7]. Nowadays commoditization and globalization are dominant themes having a major impact on business execution. As a result, major companies are focusing extensively on innovation as a significant driver of the new ideas necessary to remain competitive in this evolving business climate. Of course, 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?

With the wide use of worldwide web, people are provided a much more convenient and quick means for communication so that a much larger and richer "virtual" social networks are formed, such as "MySpace", "Facebook" and "LinkedIn". One type of the common virtual world is the forum, where people can discuss the topics of interest online at any time and any place. Therefore one approach, recently introduced by IBM, is to host an online information forum or "Innovation Jam" [5, 8] where employees (and, in some cases, external participants) are encouraged to share their ideas on pre-selected 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. Topic identification [3] poses a significant challenge in any unstructured forum like a blog, but it is perhaps less of an issue in the Jam data due to the apriori topic suggestions. As described in the following subsections, the Jam consisted of two successive phases, followed by a selection of highly promising ideas based on the 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 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 information like the organizational diversity of the participants in such threads.

In the remainder of this section, we describe the conduct of the Innovation Jam, followed by a discussion in Section 2 of the broad machine-learning challenges inherent in the analysis. Section 3 summarizes the available data and the dynamic development of the social Jam network, and Sections 4 and 5 describe respectively the unsupervised and supervised learning approaches we have applied to this data.

1.1 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 [4]. 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 implemen-

KDD'07, August 12 – 15, 2007, San Jose, California, USA.

tation details. Four general areas, called "Forums," were identified:

- **Going Places** - Transforming travel, transportation, recreation and entertainment, co-moderated by an IBM Fellow and VP of the Almaden Research Center, and the General Manager of IBM Greater China
- **Finance & Commerce** - The changing nature of global business and commerce, co-moderated by the General Manager of IBM's Managed Business Process Services, and the Global Managing Partner from IBM's Financial Services Sector, Global Business Services
- **Staying Healthy** - The science and business of well-being, co-moderated by IBM's General Manager for the Healthcare & Life Sciences Industry and the GM of Infrastructure Management Services, Global Technology Services
- **A Better Planet** - Balancing economic and environmental priorities, also co-moderated by two IBM General Managers: the GM for IBM Spain, Portugal, Greece, Israel and Turkey, and the GM for Technology Collaboration Solutions, IBM Systems & Technology

Factors that determined the selection of seed areas included: the IBM's Global Innovation Outlook (GIO), the opinions of thought leaders and technical executives, IBM's business and technical relevance, and general societal and global economic relevance.

1.2 The Innovation Jam Process

IBM's Innovation Jam was designed to take part over two phases. Phase 1 took place July 24-27, 2007 and primarily focused on ideation and development. Unlike previous IBM Jams where preparation was not necessary, the Jam required familiarization with emerging technologies which were described in on line 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 an to questions poised under a moderated topic area, forming groups which were clearly identifiable "threads" (see Figure 1).

For five weeks following Phase 1 of the Innovation Jam, a multi-discipline, international cross-IBM team, led by the IBM VP of Industry Solutions and Emerging Business, 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, 2007, and involving more focused sessions where participants refined ideas.

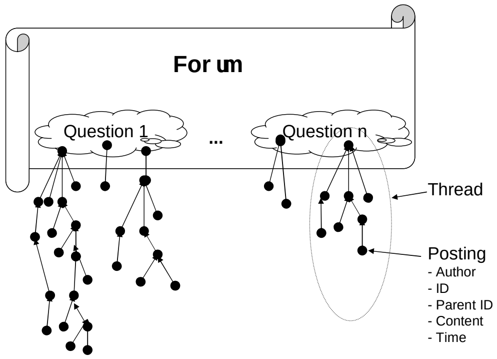

Figure 1: Relationship between postings, threads, questions and forums in both Jam phases.

After teams digested the Jam contributions, the idea finalists were selected based on follow on Market intelligence and a set of indiviso with diverse subject matter expertise to help integrate and shape some possible new innovations. Jam "finalists" were those topics identified to receive funding for development over the next two years. The ten finalists include:

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.

In the next sections, we turn to detailed description of the Innovation Jam data, as well as complementing data and our analysis. While recognizing that significant human processing took place in the course of evaluating Jam data, our goal was to see if we could identify factors that would have been predictive of the Jam finalists, perhaps suggesting ways to help make processes for future Jams less manually intensive.

2. JAM DOMAIN AND CHALLENGES

The high-level challenge of our analysis of the Innovation Jam data is to identify what are the keys to success of such an endeavor, in particular, what are the characteristics of discussion threads that lead to innovative and promising ideas. As we can see, 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, and also 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 below the results of unsupervised learning on Jam text features, which seem promising.

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 next 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, etc.

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) etc.

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 organizational and hierarchical relationships between the contributors 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 the IBM corporation, 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 — this data is of particular interest in our analysis.

3. DATA CHARACTERISTICS

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

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

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 number 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 3 times is 18% for Phase 1 and 16% for Phase 2.

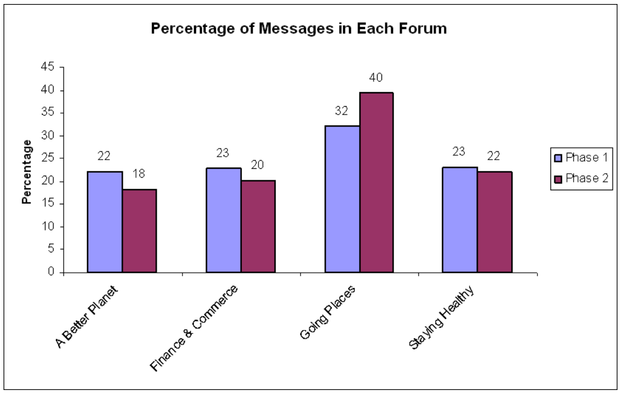

Figure 2: Percentage of messages in each forum for Phase 1 and Phase 2.

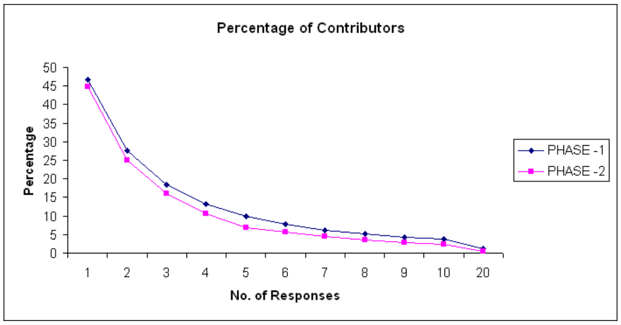

Figure 3: Percentage of contributors who responded more than 1-20 times during Phase 1 and Phase 2.

3.1 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 after an initial spike within the first two hours clear seasonality of a 24 hour rhythm. The hourly count of contributions remains fairly stable over the 3 day period.

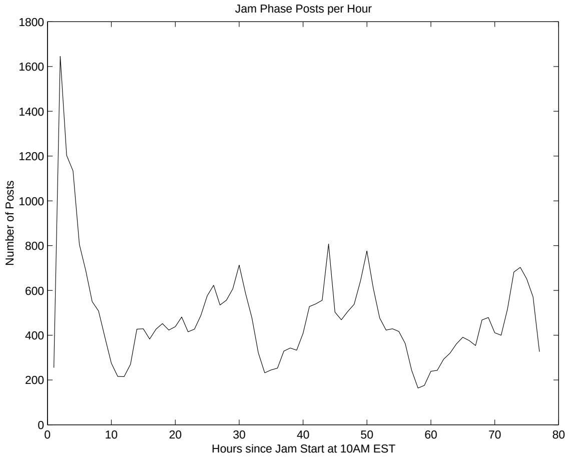

Figure 4: Number of postings over time during Jam Phase 1.

However, the cumulative number of the contributors is almost linearly increasing in time as shown in Figure 5. In the sequel we will consider the social network of contributors where every node is a contributor and a directed link 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. 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 6. The initial network has after 2 hours

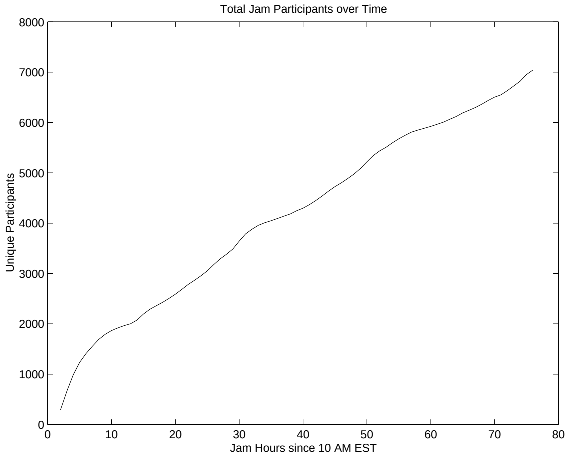

Figure 5: Cumulative number of Jam contributors over time during the first Jam Phase 1.

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.

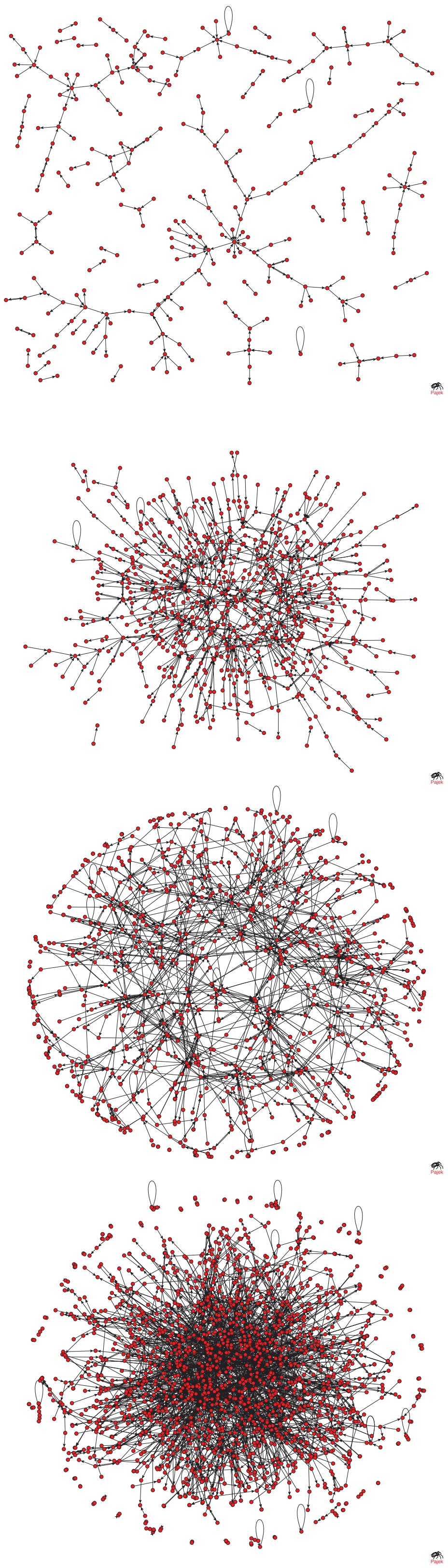

Figure 6: Evolvement of the social Jam network for 2, 3, 4 and 10 hours after the Phase 1 start.

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 interconnectivity. As a first consideration we estimate the average probability of a repeating contributor to post to a new thread, where new is defined as a thread he has never posed to. And indeed, this probability is surprisingly high at 62%. The histogram in Figure 7 shows that a large number of Jam contributors ventures 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 reveals that there is no clear functional dependence between the number of postings of a contributor and his propensity of spanning multiple threads.

4. UNSUPERVISED ANALYSIS

4.1 Preprocessing

To make unsupervised analysis of the jam data, 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 [2].

4.2 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 [6]. 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

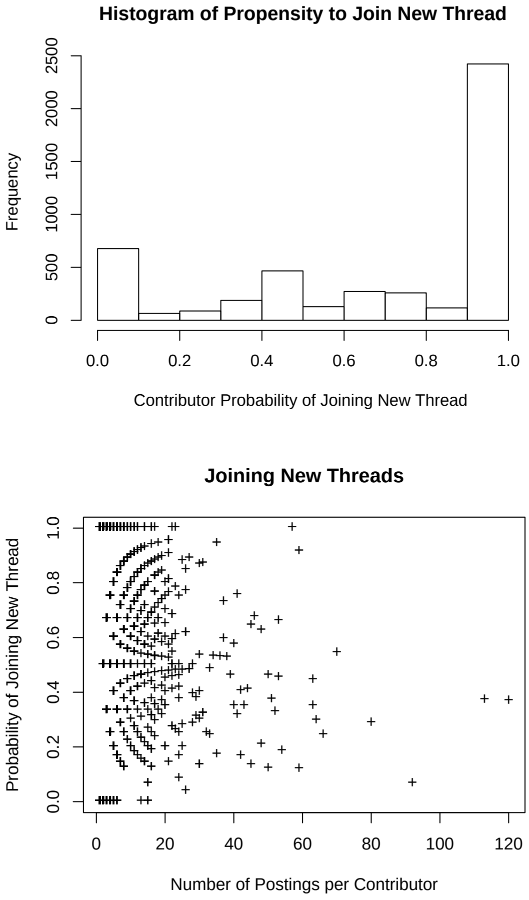

Figure 7: Histogram and scatterplot of the propensity of repeated contributors to join a new thread rather than positing to a thread they have posted to in the past.

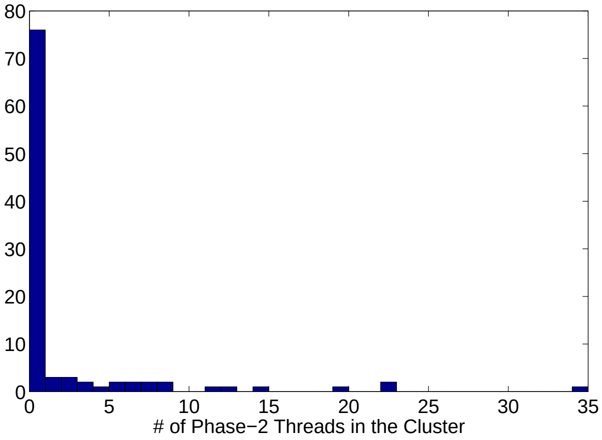

Figure 8: Histogram of the number of phase2 threads in the 100 clusters

use the open source software CLUTO [1].

4.3 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 8. 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.

5. FINDING GREAT IDEAS: SUPERVISED ANALYSIS

Several features are extracted from the Jam data. More emphasis is given to the phase 2 interactions because of the fact that the *finalists* were selected from Phase 2 of the In-

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

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.

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, flickr
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, middlewar
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, drought
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

novation Jam. A total of eighteen features (three different categories) were obtained:

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 waited 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 (containing 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 [2]. 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

Table 4: Summary of 31 big idea names obtained from the analysis of Phase 2 in the innovation Jam 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

[2] 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

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

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 [3].	0.96	0.69	0.82
T7	Standard deviation of the number of messages for all questions [3] .	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 [4].	0.31	0.70	0.34
C2	Standard deviation of the pairwise scores between the threads [4].	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 [5].	0.62	0.66	0.54
O2	Standard deviation of the pairwise distances between the contributors [5].	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

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, ref) 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. 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.

6. CONCLUSION

In this paper we have described our early efforts in analyzing the IBM Innovation Jam data in 2006. We have demonstrated how the richness of the data and its multi-faceted nature can accommodate multiple modeling approaches, trying to capture the essence of innovation and the keys to success of a Jam. As challenging the task is, our attempts have led to several interesting observations although far from reaching our ambitious goals. The contributions of our paper include: 1) the methodologies used in supervised and unsupervised analysis are directly applicable to study other forum data; 2) many of our observations and statistics from the JAM data agree well with previous work in the analysis of forum data as well as other applications, hinting that there might be universal behavioral observations held for either concentrated and controlled discussion as the Innovation Jam or the less controlled forums, such as Yahoo and Google. 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.

7. REFERENCES

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

[2] 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.

[3] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins.

Information diffusion through blogspace. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 491–501, New York, NY, USA, 2004. ACM Press.

[4] C. Halverson, J. Newswanger, T. Erickson, T. Wolf, W. A. Kellogg, M. Laff, and P. Malkin. World jam: Supporting talk among 50,000+. ECSCW'2001: European Conference on Computer-Supported Cooperative Work, Bonn, Germany, 2001.

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

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

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

[8] C. Lasser. Discovering innovation. In *IEEE International Conference on e-Business Engineering, 2006. ICEBE '06.* IEEE, 2006.

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