# Tracing Knowledge Evolution in Online Forums

Joshua Introne
MIT Center for Collective Intelligence
Cambridge, MA
jintrone@mit.edu

Sean Goggins
Drexel University
Philadelphia, PA
outdoors@acm.org

## ABSTRACT

In this article, we describe our experiences with an approach for tracing the evolution of knowledge in online conversation data. Knowledge Evolution Analysis (KEvA) works by tracking evolving clusters of co-occurring words, and reveals how knowledge flows across discussion threads and is combined by people whose activities spans threads.

We briefly present a version of the KEvA algorithm and describe its application to two different corpuses. First, we describe an analysis of small decision-making teams using two versions of an online decision support platform. KEvA identifies where participants jointly create new insights and reveals how the platform itself influences the creation of these insights.

We then describe our current efforts to scale KEvA for the analysis of a large (300,000+ messages) online forum. We describe challenges of scalability and propose approaches for overcoming them.

## Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces – *computer supported cooperative work, Evaluation/methodology, Web-based interaction, Theory and models*

## General Terms

Algorithms, Measurement, Human Factors, Measurement

## Keywords

Text mining, social network analysis, group informatics.

## 1. INTRODUCTION

When people interact online they exchange information, but digital communication is not merely transmission of bits – people manipulate and recombine information as it flows through discourse to jointly weave a shifting tapestry of ideas, perspectives, and insights.

Online forums provide a uniquely available digital trace of human interactions from which we may reconstruct these higher-level community knowledge processes. Such analysis can help to distinguish between communities, and shed light on how technology and organizational structures work together to yield networks with different characteristics and capabilities.

There is a growing interest in studying these dynamics in technologically mediated networks. For example, social tagging is a user affordance provided by some social networking platforms that researchers can use to observe how information flows across a network. Analysis of social tagging helps to illustrate the miscibility of different ideas and diffusion patterns of information [3,13].

A system user generally experiences social tagging as an explicit labeling act; the user must think to do it, do so consistently, and a critical mass of users applying tags is necessary for a tagging approach to have sustainable utility. Another user activity that can enable the analysis of information flows occurs frequently when web users copy and paste snippets of text to manage conversation flow. A side effect of these user acts is the generation of "memes" that researchers can track across multiple networks using lexical similarity algorithms [8]. Unlike social tags, lexical memes have internal structure that can be analyzed, allowing researchers to observe how these memes are transformed across network – mutating, splitting, and even merging [14].

Attempts to advance this line of research beyond memes and tags include applications of information retrieval (IR) techniques to find words or groups of words to predict other aspects of networks (e.g. [5]). These approaches offer a more general means for analyzing the flow of information in social networks, but provide little insight regarding how information evolves. IR techniques focus on classification of unstructured text, and frequently rely on the understanding that some a priori, but unobserved set of discrete topics will explain the probability distribution of words in a particular document (e.g. [2]). Such assumptions are not valid in the continual improvisation of informal conversation that is recorded in digital traces on the web.

To help study the movement of community knowledge processes, we've developed an approach called Knowledge Evolution Analysis (KEvA). KEvA is designed to be general enough to be applied to any form or online conversation, but also help to reveal the evolution and convergence of ideas in online communities. Our inspiration for the approach is a model of community evolution in the social networking literature [10], which we apply to word co-occurrence networks that are extracted from the trace of an online conversation. The approach can be used to observe how clusters of words evolve, merge, and split over time, and can be combined with other measures of group dynamics to help characterize interesting events.

In the following, we first describe the KEvA approach, and then present its application to a dataset generated during an experiment in computer mediated small group decision-making. This analysis helps to illustrate the different ways teams brought together their knowledge, and helps reveal an otherwise hidden impact of the decision support technology that was used in the study.


*WebSci 2012*, June 22–24, 2012, Evanston, Illinois, USA.

We also describe our current efforts to apply KEvA to a much larger dataset, extracted from an active community message board covering a three-year period. The size of the dataset presents many challenges, enabling us to gain insight into the specific scalability issues associated with models and algorithms studying community knowledge processes.

## 2. KNOWLEDGE EVOLUTION ANALYSIS (KEvA)

KEvA is based directly upon a technique that describes how communities of people evolve, converge, and split over time [10]. However, instead of identifying communities of people, KEvA seeks to identify "communities" of words.

The procedure works by identifying communities in a network at a series of regular time steps, and then describing how these communities change, merge, and split as the network evolves. The algorithm used to detect communities is called the Clique Percolation Method (CPM; [11].

The CPM algorithm works by first identifying a clique (a fully connected set of nodes) of some pre-determined starting size, and then replacing a node in this clique to obtain a new, overlapping clique of the same size. This process continues with each new clique, until no new replacement nodes can be found. This is euphamistically described as "walking" the clique. When a clique can be walked no further, an edge of a community has been found. A complete set of communities can be found by exhsutively applying this process to a network.

A community is defined as a set of links, so individual nodes may belong to more than one community. Once communities in each time-window have been extracted, a mapping between communities in adjacent time-windows is sought based on their similarity (see supplementary materials in Palla et al. (2007) for a detailed description of the mapping step).

To apply this algorithm to conversational data, it is necessary to transform a sequence of posts into time-sliced network data. To develop this data, transcripts are first broken into windows that contain sequences of replies. If the reply structure is not available, an initial pre-processing step in order to extract reply sequences from the conversation is of great value. The choice of windowing method is empirically determined based on the average traffic within the corpus. Our experiences have shown that it is generally useful to allow for some overlap between windows, as this helps the algorithm to construct larger communities that are a better match for the conversational topics that appear. We have also found that windows containing more than about 1500 tokens can create complexity problems for the underlying CPM algorithm (because of a critical point in underlying clique structure of the network[12]).

Data within each time window is transformed into network data using a technique called Wordij [4]. In Wordij, strong ties link adjacent words in a time window, weaker ties link words that are adjacent to a common word but are not themselves adjacent, and so forth. This approach may be applied up to some maximum degree of indirection. Within a single time window, Danowski's approach is applied separately to each sequence of replies, and the resultant networks are merged, adding edge-weights where edges overlap.

The above procedure yields a single network for each time window, and these are used as input to the community evolution algorithm. Unlike communities of people, communities of words (topics) might disappear for some time, only to re-emerge at a later point when collaborators recall them. To handle this, we keep track of all distinct, active topics and attempt to establish a mapping between these and the current time window.

The output of the algorithm is a set of topics that exist for some length of time during the conversation, and a set of links that indicate how different topics are connected. As with Palla's algorithm, topics evolve, merge, and split (see Figure 2).

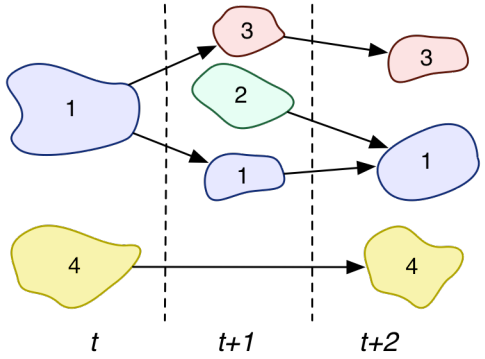



**Figure 1. Illustrating the types of relationships between topics at different time steps. Topics can spawn new children (1➔3), be consumed (1➔2), evolve (3), or disappear and then reappear (4).**

A final step in the analysis maps the actual chat traces back into the topics present in each time window. This is a simple matter of using the networks originally developed for each window and each thread, and finding the best match (in terms of relative overlap) with the extracted topics. At most one topic for each segment of chat is chosen. It is usually the case that some segments cannot be assigned to a topic, and that some topics have no posts.

## 3. STUDY 1: GROUP DECISION MAKING

KEvA was used to analyze data from an experiment exploring the impact of a decision support tool on group decision-making [7]. In the experiment, teams of five attempted to solve a murder mystery using a decision support platform. The platform was similar to a threaded chat forum, with the following constraints:

- The first post in a thread was required to be for one of three possible decision options (the suspects in the mystery).
- Users were required to indicate whether their replies agreed or disagreed with those to which they were responding.
- Users could vote on each other's posts.

In the experiment, users were required to form a consensus by the end of a time-limited decision-making period, or else forfeit an opportunity to win a small prize for making the correct decision.

The experiment compared two conditions. In the *non-mediated* condition of the platform, the platform simply provided the structure described above and a simple tool for negotiating consensus. In the *mediated* version of the platform, the system used a belief aggregation procedure to provide users with continuous feedback about which decision option was winning, and a team's final decision was constrained to match the system's assessment. If a team did not agree with the system, they could continue to deliberate and try to change the systems assessment. Thus, in the mediated condition, the platform itself became effectively an active partner in the decision making process.

The experiment was run with twenty groups of five users each. On average, the authored roughly 64 posts, and the entire corpus is roughly 14k words. The main finding reported in [7] was that the mediated groups made decisions that were more consistent with information that they exchanged, but that there was no difference in performance—about half of the groups in either condition solved the mystery.

However, qualitative analysis suggested that groups in either condition appeared to solve the mystery in very different ways. In the non-mediated condition, groups appeared to establish their solution during brief passages of intense collaborative activity during which members combined their individual pieces of information to construct a comprehensive story. This process did not appear to occur in the mediated condition.

To help quantify these differences, we sought to correlate portions of conversation where the group appeared to be merging many sources of information with a measurement of collaborative intensity. KEvA was used to identify regions of convergence conversation, and we developed a metric to help measure collaborative intensity.

## 3.1 Integrated Collaborative Intensity

In the non-mediated groups, problem solving often occurred in highly collaborative regions of the conversation. These passages had three observable properties:

1. Posting activity among a team of collaborators is focused in one thread,
2. Posting speed increased, and
3. Most of the users were directly involved in the conversation.

We developed a metric we refer to as *integrated collaborative intensity* (ICI), intended to quantify these observations. The first observation ($\theta_w$) is the inverse of the number of active threads (a thread is considered to be active in a time window $w$ if there is a post in that thread in that time window) normalized by the maximum number of threads in any time window.

The speed of posting ($\alpha_{tw}$) is determined by the number of posts in topic $t$ at time window $w$, normalized by the maximum number of posts in any window for any topic in the conversation.

Finally, the third observation ($\mu_{tw}$) is a measure of *communication integration* [1] Intuitively, it is a measure of how balanced the conversation is between team members. The value is at a minimum where no team member talks to another, and at a maximum where everyone speaks directly to everyone else. Communication integration is derived from the reply graph (an unweighted, undirected graph where each link represents a reply in chat) of users discussing topic $t$ in time window $w$, and is the average length of the shortest path through this graph between each pair of users in the team.

More precisely, let $l_{ij}$ be the smallest number of links between team members $i$ and $j$ in the reply graph for the section of conversation that is under consideration. The longest possible chain between any two members in a team of N members is N-1; if no chain exists between two members, we set the path length to N. Thus, following [1], communication integration is defined as:

$$\mu_{tw} = 1 - \frac{1}{N^2} \sum_{i \in \{1,2...N\}} \sum_{j \neq i} \frac{l_{ij}}{N-1}$$

Because each of the constituent measurements is normalized to the interval [0-1], we express ICI for topic $t$ in window $w$ simply as:

$$ICI_{tw} = \theta_{tw} \alpha_w \mu_{tw}$$

**Table 1: Statistical analysis of degree of ICI and topic convergence across the two conditions. * = p<.05; **=p<.001**

| | Mediated (*n=989*) | Non-mediated (*n=921*) |
|---|---|---|
| **Collaborative Intensity** | .032 | .038 (*) |
| **Topic Convergence** | .15 | .11 (**) |
| **Correlation** | .21 | .46 (**) |

## 3.2 Results

We applied the KEvA and ICI analyses to the collected dataset. Figure 3 provides a visualization of the algorithms output. Time flows from left to right in the graph, and each vertical line marks a minute of conversation. Each node is a segment of conversation about a topic detected by KEvA, and a node's size is proportional to ICI. Each color represents a different topic. Successive nodes from the same topic remain at the same vertical position in the graph until they merge with other topics Links indicate how topics evolve and become merged together.

We sought to develop support for the hypothesis that passages of high ICI correlated with the convergence of many different clusters of knowledge in the non-mediated condition. We measure

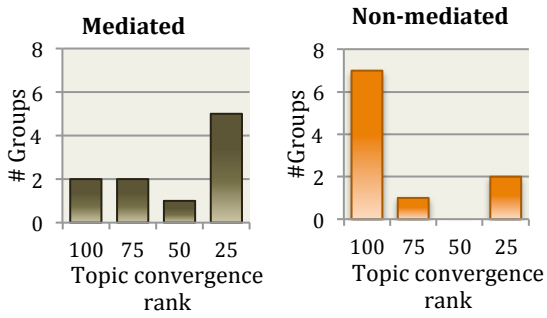
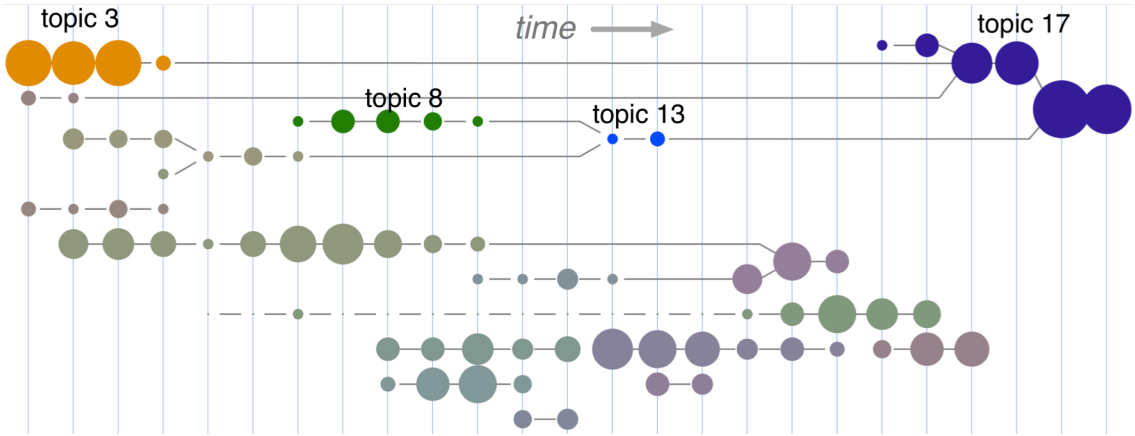



**Figure 2: Topic convergence at points of maximum ICI. Rank index on x-axis indicates upper bound of bin.**

topic convergence using KEvA as the number of distinct topics that can be connected by a path to any given passage of text. In Figure 3, topic convergence is equivalent to the number of topics present in the spanning tree rooted at each node.

We found that that the non-mediated groups exhibited less overall topic convergence, but slightly more collaborative intensity (see Table 1). A simple t-test suggested that these differences were not due to chance, but are nonetheless small. However, there was a substantial difference in the correlation between ICI and topic convergence for the two conditions, and a z-test revealed this difference to be highly significant ($Z=6.25$; $p<.001$).

To reinforce these findings, we restricted the analysis to just those periods of maximum ICI for all groups and compared the percentile rank topic convergence at these points. A histogram




| Topic 3 | |
|---|---|
| P1 | billy was VERY withholding of information |
| P2 | Has lied to police and was caught twice. |
| P1 | his fingerprints were on the crowboar, but he denied using it |
| P3 | he's just a stupid scared kid |
| P1 | and it's not his crowbar |
| P1 | then why had he handled the supposed murder weapon? |
| P4 | his fingerprints were on eddie's crowbar that was found in the bushes |
| P1 | He handled it to get to the mower - claims he never used it, though. |

| Topic 8 | |
|---|---|
| P5 | does his car have a loud muffler? |
| P3 | mrs blake didn't say anything about a loud muffler |
| P4 | if his car was there also there would have been two cars, they only heard one noise |

| Topic 13 | |
|---|---|
| P1 | eddie said he heard the loud muffler, and said it was billie's. was he trying to frame him? |

| Topic 17 | |
|---|---|
| P5 | very much disregarded the question did he find the crowbar |
| P1 | how would billie's fingerprints get on eddie's crowbar? |
| P2 | Billy said he moved the crowbar to get to the mower. |
| P1 | and would that really require throwing the murderweapo- crowbar into bushes, to be hidden? |
| P3 | no, but eddie might of used the opportunity to frame him |
| P2 | He never said he threw it into the bushes, just that he moved it - Eddit could move it |
| P2 | Eddie also wants to pin this on Billy from the beginning. (muffler) |
| P1 | what's eddie's motivation to frame billy? |
| P2 | It's somebody that is not Eddie. |

*Figure 3: Correspondence between visualization from a group in the non-mediated condition and the conversation it describes. Time flows from left to right in the graph, and passages of chat that correspond to the labeled nodes are shown beneath the graph. Links on the graph illustrate how topics become merged together in time. The size of each node reflects ICI in that time window.*

analysis of this data is shown in Figure 2. For mediated groups, periods of maximal ICI were actually better correlated with topics that had relatively low topic convergence. Precisely the opposite was true for the non-mediated groups.

These results confirmed our hypothesis, and helped to illustrate that while the groups did not differ in terms of performance on the domain task, they differed in significant way in how they processed their collaboratively held knowledge. Groups using technology that fed information about the currently "winning" decision option back to the users changed their behaviors in manner that led them to be less interested in combining their disparate pieces of information.

## 4. STUDY 2: COMMUNITY WEB FORUM

To address issues of scalability, we are now applying KEvA to a dataset extracted from the main online discussion board for the American Adult Kickball league [redacted]. Adult recreational leagues like adult Kickball serve different purposes than the same sport played as part of physical education classes; though the memory of childhood games serves as a launching pad for the sense of community these leagues engender in participants.

Imagine, if you will, US 20 something's reliving their fourth grade experience in the spirit of a college party. The combined online and offline experiences of participants result in the development of novel patterns of knowledge sharing, atypical online perspectives and information drawn from the full repertoire of their life experience to date. Specifically, the game is from childhood, but the online discourse is decorated with an obscure, specialized language for politely (or not) alluding to drinking and sexuality[9]. The league exists both in the physical world, and through an online discussion forum, which, as of 2012, has over 500,000 discrete posts from over 2,000 participants.

The specific corpus we analyze with KEvA spans 42 months and roughly 337,000 posts. Moving KEvA from a relatively small dataset of roughly 1400 posts of highly focused task-oriented users creates a new set of challenges, which we are meeting through a synthesis of KEvA with methods derived from Group Informatics [6], which is an ontology and methodological approach for modeling and transforming electronic trace data into network representations of the social experiences of online forum, social media and general collaboration system users. .

The largest hurdle is the runtime complexity of the underlying clique percolation algorithm. As described in [12], there is a threshold at which *k*-cliques merge into giant component. This point is problematic for the clique percolation method for two reasons – it obscures community structure, and it can lead to very long runtimes. For individual networks, the algorithm authors suggest a set of procedures for empirically determining the best parameterization for CPM, including setting cutoff values for weighted networks and examining the output at various clique sizes.

The community evolution algorithm involves a series of networks, and so parameters must be chosen that are effective across the entire series. Palla, et al. [10] suggest that using a single setting is effective in the case of social group analysis. However we have found that some steps in the procedure (in particular, the merging of adjacent networks in order to establish a mapping between them) can lead to substantial variance in the average degree of the network, and consequently render some parameterizations intractable for certain time-windows.

However, because networks will be partially overlapping from time-step to time-step, it is possible to maintain the structure from previous steps to reduce the complexity of subsequent, and critically, intermediate steps. This is a modest improvement upon the algorithm described in [10], but should dramatically reduce runtime complexity. We are currently developing software that implements this modification.

Another challenge is in determining the right set of parameters for use in constructing initial co-occurrence networks. Danowski[4] recommends using pairs of words as the unit of analysis building links for words up to three words away. However, this is an empirically motivated decision. It is an open research-question as to what the appropriate degree of indirection is, and what is most meaningful in online conversation.

A final (but perhaps not **the** final) hurdle in applying the algorithm to a large corpus and interpreting its results is the wide variety of purposes the message forum serves. For instance, one user in the kickball forums devotes a large number of posts (each quite long) to compose poems about the other players. Occasionally, others comment on these, but only briefly to acknowledge the accuracy or humor in a particular contribution. Members also use the forum to discuss concrete plans for social gatherings during tournaments. It is not clear at this point if the KEvA approach will offer any useful insights about such passages, or if they should be excluded from analysis.

# 5. CONCLUSIONS

We have introduced the KEvA procedure, which is a novel approach to extracting and analyzing the movement of word clusters in online conversation. Although there remains work to be done to scale KEvA to large datasets, we are pleased with its performance on smaller corpuses and anticipate that we will be able to apply the procedure to larger datasets in the near future.

# 6. REFERENCES


1. Van Alstyne, M. and Brynjolfsson, E. Global Village or Cyber-Balkans? Modeling and Measuring the Integration of Electronic Communities. *Management Science 51*, 6 (2005), 851 –868.
2. Blei, D.M. and Lafferty, J.D. Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning*, ACM (2006), 113–120.
3. Conover, M.D., Ratkiewicz, J., Francisco, M., Gonc, B., Flammini, A., and Menczer, F. Political Polarization on Twitter. *Networks 133*, 26 (2011), 89–96.
4. Danowski, J. WORDij: A word-pair approach to information retrieval. *NIST special publication*, 500207 (1993), 131–136.
5. Gloor, P.A., Krauss, J., Nann, S., Fischbach, K., and Schoder, D. Web Science 2.0: Identifying Trends through Semantic Social Network Analysis. *Computational Science and Engineering, IEEE International Conference on*, IEEE Computer Society (2009), 215–222.
6. Goggins, S., Mascaro, C., and Valetto, G. Group Informatices: A Methodological Approach and Ontology for Understanding Socio-Technical Groups. *Journal of the American Society for Information Science and Technology*, (2012).
7. Introne, J.E. Supporting group decisions by mediating deliberation to improve information pooling. *Proceedings of the ACM 2009 international conference on Supporting group work*, (2009), 189–198.
8. Leskovec, J., Backstrom, L., and Kleinberg, J. Meme-tracking and the dynamics of the news cycle. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM (2009), 497–506.
9. Novak, A. and Mascaro, C. Ti-squaring around: An Analysis of E-Jargon in an Online Sporting Community. *Popular Culture Association/American Culture Association Conference*, (2012).
10. Palla, G., Barabasi, A.-L., and Vicsek, T. Quantifying social group evolution. *Nature 446*, 7136 (2007), 664–667.
11. Palla, G., Derenyi, I., Farkas, I., and Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature 435*, 7043 (2005), 814–818.
12. Palla, G., Derényi, I., and Vicsek, T. The Critical Point of k-Clique Percolation in the Erdős–Rényi Graph. *Journal of Statistical Physics 128*, 1 (2007), 219–227.
13. Ratkiewicz, J., Conover, M., Meiss, M., et al. Truthy: mapping the spread of astroturf in microblog streams. *Proceedings of the 20th international conference companion on World wide web*, ACM (2011), 249–252.
14. Simmons, M.P., Adamic, L.A., and Adar, E. Memes Online: Extracted, Subtracted, Injected, and Recollected. *ICWSM 2011*, (2011).