iLink: Search and Routing in Social Networks

Jeffrey Davitz Jiye Yu Sugato Basu David Gutelius Alexandra Harris

Artificial Intelligence Center
SRI International, Menlo Park, CA
{davitz,yu,basu,gutelius,zaz}@ai.sri.com

ABSTRACT

The growth of Web 2.0 and fundamental theoretical breakthroughs have led to an avalanche of interest in social networks. This paper focuses on the problem of modeling how social networks accomplish tasks through peer production style collaboration. We propose a general interaction model for the underlying social networks and then a specific model (iLink) for social search and message routing. A key contribution here is the development of a general learning framework for making such online peer production systems work at scale. The iLink model has been used to develop a system for FAQ generation in a social network (FAQtory), and experience with its application in the context of a full-scale learning-driven workflow application (CALO) is reported. We also discuss methods of adapting iLink technology for use in military knowledge sharing portals and a other message routing systems. Finally, the paper shows the connection of iLink to SQM, a theoretical model for social search that is a generalization of Markov Decision Processes and the popular Pagerank model.

1. INTRODUCTION

Over the last decade, there has been an explosion of interest in social networks, stimulated in large measure by the astonishing emergence of the web as a medium for human expression and interaction. Traffic on social networking sites is rapidly outstripping traffic on more conventional web sites. To some extent the typical statistics that indicate this do not adequately capture the magnitude of the phenomenon. If one takes into account user time (in terms of both energy and focus), the social web is becoming an overwhelmingly dominant medium. The activity in social networks in terms of these other measures is beginning to exceed all other web-based activities including, for example, Internet search.[1]

A key trend is the growing importance of the social web as a production mechanism. The social web provides much more than an opportunity for people to interact and exchange general information. It is a new medium for powerful models of organizing purposeful social activities. This is compellingly illustrated in the growth of open source efforts (e.g., LAMP,[2] Wikipedia), which some authors [8, 14, 20, 27, 29] argue represent an alternate mode of social and economic production.

Much of the research in social networks has not formally modeled how social networks accomplish tasks. Rather, the work has concentrated on the structural representation, analysis, and interpretation of social network data [26, 28]. The purpose of the work discussed here is to introduce a general approach to modeling how real-time, dynamic social networks communicate and cooperate to solve problems, and to show how this understanding enables the development of applications that enhance and amplify the capabilities of these human networks. The range of potential applications is quite broad, including expertise identification and FAQ generation, social search, and smart RSS filtering.

Our high-level goals are twofold – we would like to understand how social networks create artifacts like Wikipedia, and use this increased understanding to build new applications that leverage the power of web-based social networks. To take these steps we are required to do the following:

1. *Develop enriched node and link models for social networks* – In our model, a social network is represented as a graph with nodes and links. We have discovered that in order to capture aspects of social networks that we are interested in, we need to allow the nodes to have associated property vectors. Some of these characteristics might be learned (like responsiveness) and some might be exogenous parameter values (like job title). Similarly, we must extend the link model to include, for example, the content of the interactions between nodes. In our scheme, the topology of the network is both probabilistic and dynamic.

2. *Develop an associated learning framework* – Since we are interested in building social network applications, we need to have some means by which we act within the network. We accomplish this by learning patterns of interaction and properties of the social network, and then optimizing this model with respect to some desired outcome. Performing such learning in social networks is challenging. First, the general problem can be decomposed into a number of interrelated learning problems. Second, the feedback that drives the learning comes from distributed, heterogeneous sources. The supervision must be integrated sensibly in order to pro-

[1] As reported by www.eMarketer.com in November, 2006

[2] Popular acronym for: Linux, Apache, MySQL, PHP

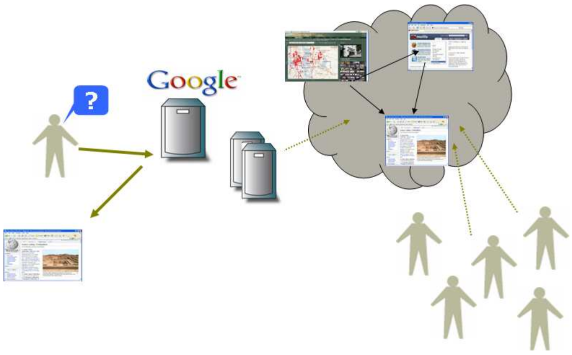


Figure 1.1: Search Engine model

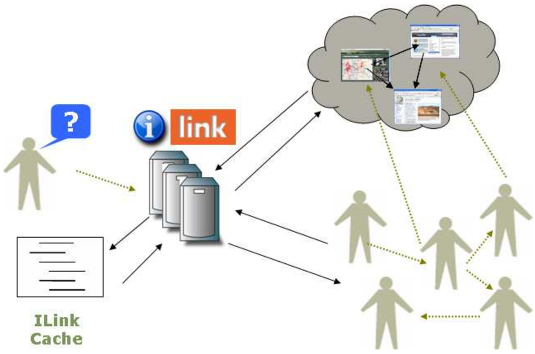


Figure 1.2: Basic iLink model

vide a useful signal to the learning algorithms.

In the remainder of this paper, we will

1. Discuss a specific model for one important kind of social production – FAQ generation. We will define the social FAQ creation model (called FAQtory), describe the architecture of an implementation and discuss some experience with the system.

2. Briefly summarize the range of research issues that we tackle currently in the FAQ generation application, and those that we will have to face as we extend the FAQtory model to a more general message routing model (called iLink) that can be applied to a fuller range of problems.

3. Show how iLink can be applied to a wide range of phenomena, by discussing real case studies of iLink deployment.

2. OVERVIEW

Tremendous effort has been expended in the development of powerful document-based information retrieval (IR) techniques. This includes traditional IR, as well as enterprise-centric and web-wide technologies. Some attention, though considerably less, has been given to the problem of identifying human sources of information, typically described as some form of "expertise identification". The extraordinary growth of the social web over the last several years has revived interest in this topic as part of a widespread surge in the study of social networks more generally.

As previously discussed, the whole notion of social network-based production has received enormous attention, no doubt motivated by hugely popular and successful open source efforts like Wikipedia and the LAMP family. Regardless of whether one sees the open source model as categorically new, it is a phenomenon of undeniable importance.

In this paper, we propose and develop an open-source style model for the production of an extremely common and widely used artifact on the web, the Frequently Asked Question or FAQ repository. This model depends on the distributed learning of expertise profiles. These learned profiles are not ends in themselves but employed as a critical component in a general model of network-based FAQ generation. That is, the goal of discovering expertise is in the service of the more general goal of constructing an FAQ repository.

This paper makes several important contributions:

1. Proposes iLink, a model for open-source content production, and introduces FAQtory, a specific application for social FAQ generation based on iLink.

2. Presents a model for distributed learning over heterogeneous sources of information in a dynamic social network.

3. Discusses extensions of the specific FAQtory model to a general message routing system.

4. Outlines connections of iLink to SQM, a recently proposed theoretical social query model [2].

We now discuss these contributions in detail.

A specific open-source model. Many discussions of open-source production systems do not attempt to model the underlying dynamics. Rather they are described informally in terms of shared practices or collaboration platforms (like Wikis). This leaves the impression that the social networks underlying the open-source production systems work as loosely or semi-organized mobs. We believe that this is an inadequate model of the underlying social phenomena. Wikipedia and Linux communities may be decentralized and distributed, but there is considerable evidence that there are, in fact, systematic processes in the activities of the social networks that account for their success [9]. Our model reflects some of the important characteristics of how social systems generate and distribute messages, something that real social networks clearly carry out, even if informally.

The iLink model and the resulting FAQtory application based on it make it possible to scale and amplify these informal processes. We propose a recognizably open-source model for the production of a highly useful artifact, the FAQ. This FAQ can be usefully thought of as a kind of Wikipedia where the content is highly specific (tied to a question) and demand-driven (in response to a query). As a result, FAQs produced by the FAQtory application share many of the positive features of the Wikipedia, in particular the fact that they are continuously generated and revised. The iLink model also provides a principled approach to an open source model of production, with potential benefits for other open source opportunities beyond FAQ generation.

Distributed learning over heterogeneous sources. In the iLink model, messages are routed by a learning system that integrates information over multiple and heterogeneous sources. Unlike many models of expertise identification and answer generation, we allow for incremental answering. At each step in a query thread in FAQtory, user nodes can contribute some information even if that information does not qualify as an answer. This information can be about the query itself or it can simply be some evidence about where knowledge might exist in the network (e.g., who knows something, who knows somebody). The learning system takes continuous feedback in order to improve how message threads get routed from any node.

The learning in iLink occurs by watching a natural social network and selecting over effective strategies surfaced by this system as the members try to resolve queries. Since the learning system is continuously monitoring the real social network, it is capable of drafting off of the social network's learning. In some fashion, iLink enables the social network to discover and amplify its own capabilities.

Extension of the specific model to a general message routing system. Even though FAQtory is a social Q/A system, there is nothing in the iLink model that requires the originating messages to be explicit questions. In fact, a system can learn by similar mechanisms as FAQtory over general message passing, since the learning can operate over general messages without much modification. We will briefly describe how the iLink model can be extended to other message routing applications (e.g., smart RSS filtering). Note that certain features in the FAQtory system (e.g., the answer validation mechanism) may become irrelevant or unnecessary in the general message routing scenario.

A general paradigm for social search. We show the connection of iLink to Social Query Models (SQM) [2], a theoretical model for social search that includes Markov Decision Processes and the popular Pagerank model as special cases.

3. OVERALL MODEL

An abstract model can be used to characterize various aspects of the iLink system, e.g., specification of properties of nodes in the network, flow mechanism during message routing. The iLink model has various components:

1. **Node**: The network has a set of n *network nodes* $N = \{N_i\}_{i=1}^{n}$. As mentioned earlier, each node represents a user in the network, and has an associated profile.[3] In this model, the profile of the i^{th} node is considered to comprise of a set of parameters: (a) an expertise measure E_i, represented by probability distributions $P(T_k|N_i)$ over a global topic set $T = \{T_k\}_{k=1}^{t}$; each topic is itself a probability distribution over identifiers, where an identifier could represent a wide variety of entities, e.g., message words, meta-tags, entries in an ontology, (b) a referral rank F_i, which measures the general affinity of other nodes to route messages to the i^{th} node and is calculated using an iterative computation over the referral links, in a way similar to the Pagerank computation, and (c) a response score R_i, which is a function of the response rate and response accuracy of a node to incoming messages. The profile information of all the network nodes is maintained at the supernode, which is defined next.

2. **Supernode**: A *supernode* S has the following components: (a) a database D storing all the past message streams, (b) profile parameters E, R, F for all the nodes in the network, (c) the set of all possible topics T.

3. **Message**: A *message* m is routed between a node and the supernode. A new message generated by any node is represented as m_0. As content or meta-content gets added to the message by each node while being routed around in the network, the modifications made to the message are represented as m_k, where m_k is the k^{th} modification made to the initial message m_0. A general message can have either content (e.g., words), meta-content (e.g., tags) or a forwarding address. We will discuss two specific types of messages: (a) a question q, and (b) an answer a. A thread of messages, starting from m_0 to the last modification m_k added to the message, constitutes a *message stream* initiated by m_0.

For ease of explanation, we will describe the detailed message flow in the FAQtory system, where the initial message m_0 in a stream is a question q, and the final message is an answer a – therefore in the following discussion, a message stream will always begin with a question and end with an answer. All message flows between nodes are moderated by the supernode S, i.e., all messages are sent via S. On getting a message, a node has the choice to (i) answer, (ii) ignore, (iii) route the message, possibly with modifications to the incoming message, or (iv) express interest in validating the answer to the query. This model has two modes of possible routing scenarios. A node can either (a) annotate a message with a forwarding address and send it to the supernode S, specifying that it be routed to another node, or (b) annotate a message with other content and forward it to S, in which case S makes the decision about to whom to next route the current message.

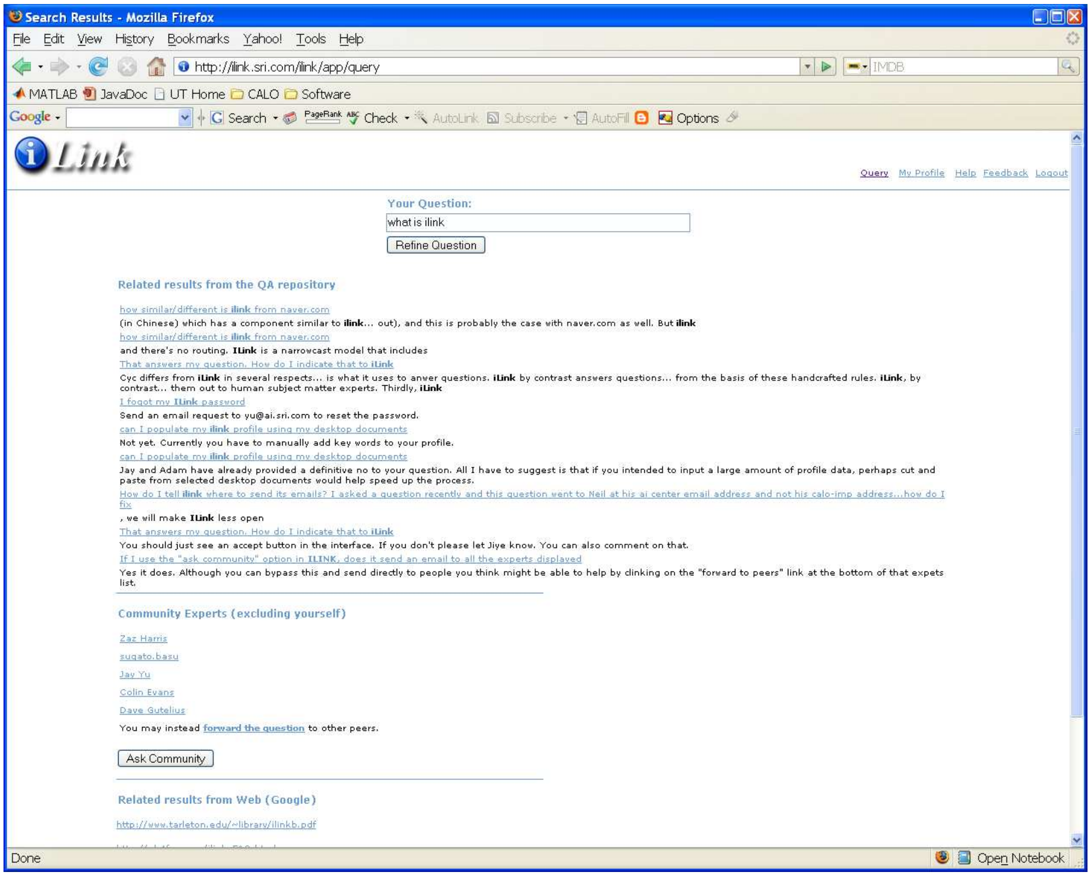
Here is how the message flow and routing mechanism works in FAQtory:

1. A new message stream $\{m_0\}$ (where $m_0 = q$, a question) is always sent to the supernode S.

2. For each incoming message stream $\{m_0, \ldots, m_k\}$, S computes its match with the existing *message streams* in the history database D. If a good enough match is found, then the answer a of the matching message stream from D is extracted and recommended to the asking node as a possible answer for q. If the asking node accepts the answer a, then the current message flow is terminated. Otherwise, it proceeds to step 3.

3. If m_k in $\{m_0, \ldots, m_k\}$ is marked as an answer a, then S routes it directly to the node that initially asked the question $q = m_0$. If the asking node confirms that a is indeed a correct answer to q, then the message flow in the system is terminated by S by notifying the nodes involved in the flow that a correct answer has been obtained. If the asking node cannot confirm the correctness of a, S simultaneously sends the answer a for validation to those nodes that had previously expressed interest in evaluating the answer during query routing. The answer is given the combined voted score of the evaluators. If a is not a correct response for q, it continues to step 4.

[3] We will use the term node to refer to peers, users or members of the social network.


Search Results - Mozilla Firefox
File Edit View History Bookmarks Yahoo! Tools Help
http://ilink.sri.com/ilink/app/query
IMDB
MATLAB JavaDoc UT Home CALO Software
Google Search PageRank Check AutoLink Subscribe AutoFill Options
iLink
Query My Profile Help Feedback Logout
Refine Question
Ask Community
Done
Open Notebook

Your Question:

what is ilink

Related results from the QA repository

how similar/different is **ilink** from naver.com

(in Chinese) which has a component similar to **ilink**... out), and this is probably the case with naver.com as well. But **ilink**

how similar/different is **ilink** from naver.com

and there's no routing. **ILink** is a narrowcast model that includes

That answers my question. How do I indicate that to **iLink**

Cyc differs from **iLink** in several respects... is what it uses to anwer questions. **iLink** by contrast answers questions... from the basis of these handcrafted rules. **iLink**, by contrast... them out to human subject matter experts. Thirdly, **iLink**

I fogot my **ILink** password

Send an email request to yu@ai.sri.com to reset the password.

can I populate my **ilink** profile using my desktop documents

Not yet. Currently you have to manually add key words to your profile.

can I populate my **ilink** profile using my desktop documents

Jay and Adam have already provided a definitive no to your question. All I have to suggest is that if you intended to input a large amount of profile data, perhaps cut and paste from selected desktop documents would help speed up the process.

How do I tell **ilink** where to send its emails? I asked a question recently and this question went to Neil at his ai center email address and not his calo-imp address...how do I fix

, we will make **ILink** less open

That answers my question. How do I indicate that to **iLink**

You should just see an accept button in the interface. If you don't please let Jiye know. You can also comment on that.

If I use the "ask community" option in **ILINK**, does it send an email to all the experts displayed

Yes it does. Although you can bypass this and send directly to people you think might be able to help by clinking on the "forward to peers" link at the bottom of that expets list.

Community Experts (excluding yourself)

Zaz Harris

sugato.basu

Jay Yu

Colin Evans

Dave Gutelius

You may instead **forward the question** to other peers.

Related results from Web (Google)

http://www.tarleton.edu/~library/ilinkb.pdf

Figure 3.3: FAQtory screenshot

4. If m_k specifies a direct forwarding address, then S routes the message to that address directly. Else, it continues to 5.

5. If m_k does not specify a direct forwarding address or an answer, then for each node N_i in the network, a ranking score is computed by S using a weighted linear combination (weights are currently an input parameter) of the node's current score F_i, response score R_i and the similarity between the topic distributions of the message and the node's expertise measure E_i, calculated as $\sum_k P(m|T_k)P(T_k|N_i)$. The nodes are ranked by S in decreasing order according to this ranking score, and the current message stream $\{m_0, \ldots, m_k\}$ is next routed to the top r nodes in the network (r is an input parameter).

6. Each of these r nodes now has the choice to ignore the message stream, add a comment or address m_{k+1} to the message stream and send it to S, or add an answer a and send it to S, after which steps 2-5 of the message flow are repeated. This message flow continues until the message stream is terminated by a correct answer a to q.

7. On a message flow getting terminated by a correct response, the profile parameters E, R, F are updated globally at the supernode for all the nodes involved in the routing, and the set of topics T is also updated if necessary. Details of these updates are discussed in Section 5.

The goal of the system is to effectively learn the E, R, F parameters of the node profiles and the global topic set T, so that an answer a is obtained for a question q that initiated the message stream by using the smallest number of routing steps. In the practical system, one can make enhancements, e.g., the referral score and the response score can be topic based. A user node may have low response rate on average, but on a specific topic (e.g., music, mathematics) may be willing to go out of his way to correctly respond to or at least tag the question. For example, let $P(R_i|T_k)$ be the probability of response given a topic. Then, for any given message m, the topic-specific response rate can be calculated as $P(R_i|m) = \sum_k P(R_i, T_k|m) = \sum_k P(R_i|T_k)P(T_k|m)$.

Note that the iLink model can in general be used for routing any messages (e.g., news articles), so in practice the message stream need not necessarily always start with a question and terminate with an answer, as it does in the FAQtory system.

4. SYSTEM ARCHITECTURE

The current implementation of the FAQtory system is a stand-alone web application that functions as the supernode in the social network, facilitating the generation of a repository of question/answer threads. FAQtory users query the system and are presented with a list of question and answer pairs that are related to the query, a list of experts on the topics found in the query, and as a last resort, search results from the web. The query can be forwarded to the experts or to other nodes not listed, and the recipients of the forwarded query may choose to answer the question, indicate interest in future answers, or request clarification. The resulting answers are rated and reviewed by the asker and interested experts. All these interactions are routed through the server hosting the web application, which functions as the supernode and estimates different model parameters, e.g., the expertise E, response rates R, and referral rates F of the nodes involved, and the topic set T. Figure 3.3 shows the FAQTORY screenshot of the result of the query "What is iLink" – it shows top matching results from the FAQ, shows a list of relevant community members to whom the question can be routed, and a list of webpages matching to the query.

At the architecture level, the FAQTORY system is implemented as a three-tiered client-server application. The client/front tier facilitates interactions between the nodes and the FAQTORY server via APIs. This allows network nodes to access FAQTORY through various channels, e.g., web browser, email clients, instant messenger. The middle tier is the core of the system and it consists of several components, as shown in Figure 4.4. The router component dispatches messages to proper handlers: (1) queries to QA handler to fetch (if any) the relevant existing answers from either the KB repository (FAQ) or the third party source such as the web search engines; (2) queries to Peer handler to fetch (if any) the relevant experts from the network/user community; (3) the server responses to Messenger to send back to the nodes.

The handlers get their answers from the Knowledge Managers (KM) that interact with the back-end knowledge bases (e.g., database, text documents, emails), primarily through an indexing engine and a ranker. The KMs also interact with a (topic) Learner, which is responsible for creating and updating the topic models for the nodes. The interactions between the KMs and the indexing engine as well as the learner are based on the event-listener design pattern, which allows incremental and dynamic updates of the index and the topic models, thereby facilitating online learning.

The analysis and processing of the free text is handled by the NLP processor that performs stemming, synonym lookup, stop word removals, and so on. The index engine is used to create inverted indices on both the existing answers and the node expertise, which allows fast retrievals by the KMs. The index of existing answers makes use of not only the keywords from the answers, but also the topics to which they belong and the answer ratings the node provided. While indexing user nodes and ranking them, FAQTORY takes into account their expertise E as measured over content (e.g., profiles, answers), their referral rank F (the popularity by referrals from other nodes), as well as the response score R (function of response history parameters, e.g., response rate, response accuracy) of the nodes.

5. LEARNING FRAMEWORK

We will now discuss the details of the learning framework that is implemented in FAQTORY, to deploy the model described in Section 3. The learning framework in this case has to solve a set of interrelated learning problems, and we have an integrated approach for doing this. The main research requirements that drove choices in such a model are listed below, with the corresponding algorithmic techniques that are currently used to address these concerns in FAQTORY. These issues are also relevant for other system deployments of the iLINK model, since none of these issues are FAQ-specific.

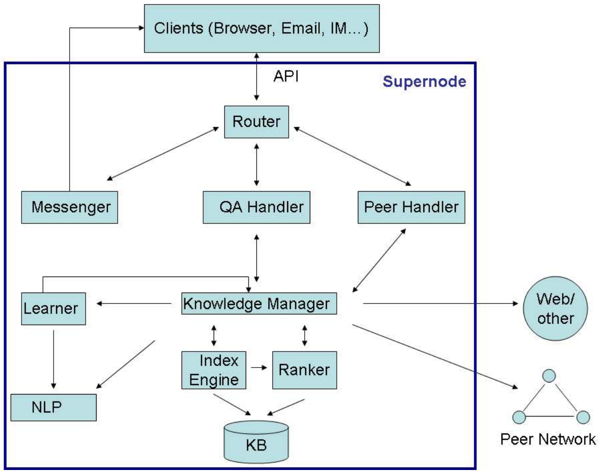


Figure 4.4: FAQTORY **system architecture**

1. **Heterogeneous topics**: As mentioned earlier, the identifiers in the topic set T have to represent diverse objects like message words and meta-tags on messages. The challenge, therefore, is to map the diverse types of objects into a common representation.

Solution: In the current system, every topic T_j in the topic set T is a probability distribution over identifiers w, where each identifier in w could be a word in a message, a tag in a meta-content annotation, a descriptor for an ontology node, and so on. The topic distributions for the different data types (e.g., words, tags) can be maintained separately, if necessary.

2. **Cold-start problem**: In some domains profile information is already available for the nodes in the network (e.g., from org chart data, webpages). However, this may not be available in all domains. In that case, iLINK should be able to infer profile distributions for nodes from multiple data sources, e.g., homepage, email, documents, org charts.

Solution: iLINK can currently learn a multirelational topic model [4] for finding latent topic distributions from multiple sources, e.g., emails and other documents composed by a user, using a scalable iterative algorithm to perform relational clustering of data from multiple heterogeneous sources. This step outputs (a) a set of local topics L_i, where the j^{th} local topic L_{ij} is a distribution over a set of local keywords, and (b) a probability distribution $P(L_{ij}|N_i)$ over the local topic set L_i for each user node N_i.

3. **Privacy issues**: While doing profile inference, a critical concern is privacy, since in most cases users will not be willing to share sensitive information, e.g., emails. In order to infer distributions of node topic preferences over a global topic set, local topic models have to be first run at each user node to get the distribution for each node over a local topic set L_i – the challenge here is to infer a global topic set T

from each of these individual local topic sets L_i, and then modify the local node topic profiles to obtain the expertise measure E_i over this global set of topics T.

Solution: The topic models on sensitive documents are learned locally at each node, so that there is no violation of privacy concerns. The outputs of the local topic models are sent to the global supernode S, which treats the j^{th} local topic L_{ij} from node N_i as a point in a manifold over identifiers and performs information-theoretic probabilistic clustering (EM using KL divergence) [13] of these topics. This can be further improved by using constrained clustering [5] for this task, where cannot-link constraints are added between topics of the same node to prevent topics of a particular node from being clustered together due to uniqueness of a node's vocabulary. Each cluster centroid is now considered as a global topic T_k (the number of global clusters t has to be input in the current formulation).

To estimate E_i, the probability distribution over topics $P(T_k|N_i)$ is obtained as follows: $P(T_k|N_i) = \sum_{L_{ij}} P(L_{ij}|N_i)$ $P(T_k|L_{ij}, N_i)$, where $P(T_k|L_{ij}, N_i)$ can be obtained from the posterior probability of local topic L_{ij} being assigned to the centroid representing T_k, while $P(L_{ij}|N_i)$ is obtained from the local topic model of the node. The final output of the clustering is the set of global topics T and the probability distributions $P(T_k|N_i)$ for each node N_i, constituting their expertise measure E_i. Note that other privacy-preservation techniques can be used here, too [24].

4. **Prior knowledge**: The topics should be able to incorporate partial supervision like domain-specific ontologies (e.g., org chart in an enterprise application, UMLS[4] in a medical application), user-specified tags (e.g., from manual profile edits), provided as prior knowledge to the system for topic inference.

Solution: A set of keywords/tags input as non-hierarchical semi-supervision is incorporated into the topic model by using an empirical prior over the relevant keywords. Currently, the hierarchy structure in hierarchical supervision is not used in learning the topic models model – hierarchical priors like ontologies are represented by considering only the leaf nodes in the hierarchy as a set of keyword identifiers, which are then incorporated into the topic model.

5. **Message matching**: The supernode S should be able to find matching relevant messages for a new message by looking up the Q/A repository D to potentially get an answer to a question without any routing.

Solution: Any time a message stream $m = \{m_0, m_1, \ldots, m_k\}$ passes through the supernode S, it finds the best p topics from T for the current message stream segment m, calculated according to the probability score $P(m|T_k)$. It then augments the keywords in the message stream with the keywords (suitably weighted by their corresponding probabilities) in the top p scoring topic distributions. Next, S computes the cosine similarities between the current augmented message stream and the existing message streams in the history database D. Finally, S returns the answer of the message stream from D that has the highest match score.

6. **Scalability**: Any learning algorithm used by S should be able to scale up to potentially a large number of network nodes and large size of the topic set T.

Solution: To address scalability concerns, pseudo-linear algorithms are used for topic inference and inverted indices are used for efficient implementation. iLink maintains two inverted indices: I_1 for mapping keywords w to topics T, and I_2 for mapping topics T to nodes N. When a new message stream m comes into the system, I_1 takes m as input and outputs a vector of weighted topics relevant for m. I_2 next takes this vector of weighted topics as input and outputs a weighted vector of relevant nodes. Both of these inverted index lookups are very efficient. The output of I_2 is used to get a ranked list of nodes relevant to message m.

7. **Incremental learning**: iLink should be able to infer new topics and accordingly update the topic set T and the node profile parameters E, R, F online, as new messages flow through the system and as new nodes get added to the network.

Solution: Two kinds of incremental updates need to be made to the parameters E, R, F and T at the supernode S — (a) whenever a message stream terminates (for FAQtory, this corresponds to a correct answer a being found for a query q), and (b) when a new node is added to the network.

In the first case, the following updates are performed:

(i) Every topic T_k in the global set of topics T can be updated incrementally as follows: $T_k^{(t+1)} = T_k^{(t)} + \frac{1}{t+1}(m - T_k^{(t)})$, where m is the corresponding message stream [3]. This way, every topic is updated according to its probability of producing the message stream.

(ii) The referral rank scores F of the nodes involved in routing the message stream are updated using an algorithm similar to incremental Pagerank update.

(iii) The response rate and response accuracy in the response score R_i are updated for each node N_i that participated in the message stream that generated the correct response. Topic specific scores and updates are used in a practical setting. We can consider $P(R_i|T_k)$ and $P(F_i|T_k)$ and do the appropriate updates.

(iv) Finally, the probabilities $P(T_k|N_i)$ in the expertise measure E_i have to be updated for every node N_i participating in the message stream. For every message m generated by the node N_i in the message stream, the posterior probability $P(T_k|N_i, m)$ on observing m is calculated using Bayes Rule: $P(T_k|N_i, m) \propto P(m|T_k, N_i).P(T_k|N_i) = P(m|T_k).P(T_k|N_i)$. After renormalization, these posterior probabilities $P(T_k|N_i, m)$ constitute the updated expertise measure E_i.

In the second case, when a new node gets added, iLink runs the topic model on the new node's documents to get local topics. The supernode can then assimilate these local topics into the global topic set T by either rerunning the clustering or assigning the new topics to the closest existing cluster centroids. In the second case, the E, R, F parameters of the other nodes are left unchanged – the E for the new node is assigned to the calculated value, while its R and F parameters are set to default values.

Note that in both cases, the number of topics remains the same and only the internal keyword distribution in each topic is updated. In each case, the inverted indices are suitably updated to reflect these changes.

[4] Unified Medical Language System

6. CONNECTIONS TO SQM

Interesting connections exist between the social query aspect of iLink and the Social Query Model (SQM), a recently proposed model for social query routing [2]. SQM is a model for decentralized search, taking into account social interactions that include actions like generating, gathering, sharing, and distributing messages by different network nodes. SQM provides a generalized social rank metric that includes the Pagerank [11] model and Markov Decision Processes [7] as special cases. SQM explicitly models realistic parameters like response rate, expertise, correctness and routing policies of nodes in a network, and proves the existence of a query routing policy that is simultaneously optimal for all nodes. SQM models query routing in a social network as a multiparty game, where the payoff of the query routing policy is the probability of each node getting an answer to a query. Given the network parameters, there is a near-optimal efficient algorithm in the SQM model for calculating this policy, which is linear in the number of network links.

SQM provides a formulation for decentralized search. But it can be easily adapted for use in the iLink formulation, which has a centralized supernode. The supernode can observe message routing between the nodes and learn maximum likelihood estimates of model parameters (e.g., response rates, expertise) from routing traces. Using these parameters, it can infer the value of getting an answer through different nodes by calculating the payoff of the SQM model, and decide which node to route a message to on getting a query by using the near-optimal SQM policy estimation algorithm.

7. CASE STUDIES AND APPLICATIONS

We next outline some case studies where iLink and its underlying learning technology have been (or could be) deployed in real applications.

7.1 Calo Test Pilot

The FAQtory system of iLink was developed as part of CALO, an adaptive cognitive system funded under DARPA's PAL (Perceptive Assistant that Learns) program. The goal of this program is to develop an integrated learning system that adapts and builds new capabilities in support of individual workflow. Social networks are essential to information workers, often significantly more important than online electronic information sources. iLink is designed to help facilitate the user's access to this resource as well as improve the functioning of the resource itself. Funding in the program is contingent on passing a yearly test of the integrated CALO system. The test involves live use of the system in real working environments. Fifteen users used the CALO system over a several-week period, exercising all its capabilities including iLink. As discussed in Section 4, the FAQtory feature of iLink was deployed as a web application and it supported routing of user questions. At the end of the test period, the users built up a FAQ repository. In building this FAQ, FAQtory also built up node profiles, a feature that could also be used for other iLink applications such as RSS filtering and group formation.

The general goals of testing iLink included demonstrating: (1) real-time learning by matching queries and community users, (2) adaptability to user demands and direction, (3) accuracy in message targeting and routing, (4) dynamic user profile correction based on community behaviors and identification of community "experts" on a given topic at a given point in time, (5) improvement in the overall quality of resulting query-answer pairs, (6) dynamic topic refinement and meta-content generation, and (7) acceptable performance under load.

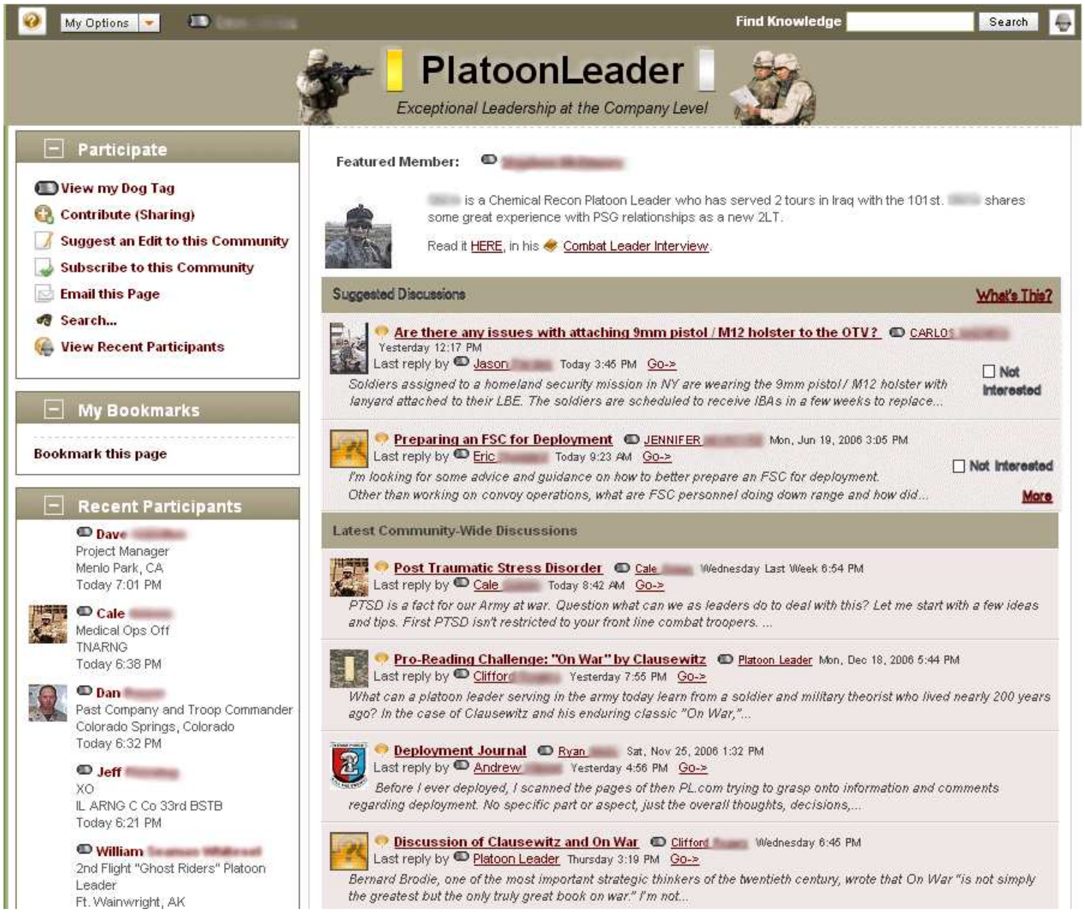
The goal of the CALO test was to evaluate the system as a whole and so there was no explicit test of the iLink system in isolation. However, the system's performance on areas that involved iLink was extremely strong, e.g. query routing. While the results were somewhat anecdotal, it was clear that the iLink system was simple to use (the user interaction model follows search and email models) and that the FAQ that was produced appeared to avoid some of the issues characteristic of other types of open contribution question answering mechanisms, e.g., bulletin boards. This seemed to be a result of the fact that iLink returns QA pairs that are embedded in a generating thread. The thread provides rich context for the question-answer pairs that helps enhance user recognition of extant answers. That is, the value of the extant answers is enhanced for users asking questions close to ones already handled through the system. We plan to test this hypothesis in detail in the military applications that will be described next.

7.2 PlatoonLeader

The first pilot deployment target for iLink is integration of the FAQtory concept into an existing online community named PlatoonLeader,[5] a professional forum for past, current, and future U.S. Army platoon leaders. PlatoonLeader is organized around several topics (e.g. Leadership, Fitness) that are led by topic leaders who facilitate discussions, develop relevant content, and encourage member participation. Upon joining, members create a profile, described as a "dog tag", participate in threaded discussions, and upload and comment on files, referred to as "knowledge objects". The initial topic profile for each user will be determined from the dog tag of the user and augmented by indexing the knowledge objects and discussion posts contributed or commented on by that user. Within a discussion thread, the initial post will be considered analogous to the FAQtory query and subsequent posts as answers to or clarification of the question. Other events or data logged by the system or otherwise accessible may also be used to inform iLink topic profiles.

PlatoonLeader will be provided with two features that leverage iLink technology: "Suggested Discussions" and the "Moderator's Assistant". Immediately upon logging in, every user will see "Suggested Discussions", that is, the most recent discussion threads considered relevant based on his or her topic profile (see Figure 7.5). The user can provide direct feedback to iLink by indicating whether the thread was a good match, or indirectly by viewing or posting to the thread. The second feature, "Moderator's Assistant", will be available only to topic leaders, who currently spend a lot of time reading all the discussions in their topic areas and manually emailing requests to knowledgeable members who are not yet participating to join the discussion. When reviewing a discussion thread, the topic leader will be presented with members that iLink has identified as experts on the thread topic. The topic leader will have the option to send an automated request to contribute to the thread to the experts suggested, or to members not identified as

[5] http://platoonleader.army.mil



My Options [redacted] Find Knowledge Search

PlatoonLeader

Exceptional Leadership at the Company Level

Participate

- View my Dog Tag
- Contribute (Sharing)
- Suggest an Edit to this Community
- Subscribe to this Community
- Email this Page
- Search...
- View Recent Participants

My Bookmarks

Bookmark this page

Recent Participants

Dave [redacted]
Project Manager
Menlo Park, CA
Today 7:01 PM

Cale [redacted]
Medical Ops Off
TNARNG
Today 6:38 PM

Dan [redacted]
Past Company and Troop Commander
Colorado Springs, Colorado
Today 6:32 PM

Jeff [redacted]
XO
IL ARNG C Co 33rd BSTB
Today 6:21 PM

William [redacted]
2nd Flight "Ghost Riders" Platoon Leader
Ft. Wainwright, AK

Featured Member: [redacted]

[redacted] is a Chemical Recon Platoon Leader who has served 2 tours in Iraq with the 101st. [redacted] shares some great experience with PSG relationships as a new 2LT.

Read it HERE, in his Combat Leader Interview.

Suggested Discussions — What's This?

Are there any issues with attaching 9mm pistol / M12 holster to the OTV? CARLO[redacted]
Yesterday 12:17 PM
Last reply by Jason [redacted] Today 3:46 PM Go->
Soldiers assigned to a homeland security mission in NY are wearing the 9mm pistol / M12 holster with lanyard attached to their LBE. The soldiers are scheduled to receive IBAs in a few weeks to replace...
□ Not Interested

Preparing an FSC for Deployment JENNIFER [redacted] Mon, Jun 19, 2006 3:05 PM
Last reply by Eric [redacted] Today 9:23 AM Go->
I'm looking for some advice and guidance on how to better prepare an FSC for deployment. Other than working on convoy operations, what are FSC personnel doing down range and how did...
□ Not Interested
More

Latest Community-Wide Discussions

Post Traumatic Stress Disorder Cale [redacted] Wednesday Last Week 6:54 PM
Last reply by Cale [redacted] Today 8:42 AM Go->
PTSD is a fact for our Army at war. Question what can we as leaders do to deal with this? Let me start with a few ideas and tips. First PTSD isn't restricted to your front line combat troopers. ...

Pro-Reading Challenge: "On War" by Clausewitz Platoon Leader Mon, Dec 18, 2006 5:44 PM
Last reply by Cliffor[redacted] Yesterday 7:55 PM Go->
What can a platoon leader serving in the army today learn from a soldier and military theorist who lived nearly 200 years ago? In the case of Clausewitz and his enduring classic "On War,"...

Deployment Journal Ryan [redacted] Sat, Nov 25, 2006 1:32 PM
Last reply by Andrew [redacted] Yesterday 4:56 PM Go->
Before I ever deployed, I scanned the pages of then PL.com trying to grasp onto information and comments regarding deployment. No specific part or aspect, just the overall thoughts, decisions,...

Discussion of Clausewitz and On War Clifford [redacted] Wednesday 6:45 PM
Last reply by Platoon Leader Thursday 3:19 PM Go->
Bernard Brodie, one of the most important strategic thinkers of the twentieth century, wrote that On War "is not simply the greatest but the only truly great book on war." I'm not...

Figure 7.5: iLink PlatoonLeader integration

experts. This is analogous to the forwarding feature implemented in the FAQtory and would similarly train the iLink supernode.

7.3 Other Messaging Applications

In the FAQtory application, it is assumed that message streams are initiated as questions. But aside from the answer that triggers a validation step, there is nothing in the model that requires threads to be initiated with a question. General messages can be sent to the iLink supernode with the implicit question of who should see this. Given such a system, for example, an FBI agent who discovered some suspicious people were taking flight lessons would have been able to simply narrowcast this discovery to a relevant set of people (selected automatically by the iLink supernode) with the attached implicit question: 'who ought to see this message?'. Here are some messaging applications where iLink could be useful:

1. **Smart RSS Filter**: Social networks are constantly faced with the problem of routing discovered information to the right consumers. The iLink model is extremely well positioned to address this fundamental problem. We are currently designing a smart RSS application that takes advantage of the profiling and routing capabilities of the system in a style similar to the FAQtory. The goal of this system is to provide a filter for each user that can filter the RSS feeds a user subscribes to and suggest messages (e.g., announcements, news articles) that the user might be interested in, based on individual preferences of the user as well as aggregate preferences of the user's social community. The goal here is to create a social RSS system that taps the abilities of the network to direct messages – the iLink supernode watches a user's routing behavior to refine the user profiles that are used to tune the filters, as well as suggest messages that the user's "friends" in the network are interested in.

2. **Message Routing for Advertising**: An important additional application of iLink includes improving click-through rates for web-based advertising. Currently, the most sophisticated ad targeting strategies focus on keyword matching of text for a given web page. iLink can add an important contextual component to such a strategy by paying attention to (1) a user's set of interests at a given point in time; (2) that user's set of friends who share certain features in common at a given point in time, as inferred from the user's social interactions. Site owners, for instance, can use this information in addition to current keyword-only-based tools to identify better-matched ads to target the users in their social network, improving cost efficiencies of ad placement and conversion rates.

8. RELATED WORK

Research in social networks covers various areas, some of which are (i) analysis of general network properties, e.g., degree distribution of nodes in the network, shortest path between nodes in small-world graphs [32]; (ii) generative models, e.g., random graph models, preferential attachment models, and their resulting statistical properties [25]; (iii) dynamical models, e.g., economic transactions, disease transmissions, spread of innovations [34]; (iv) models of temporal evolution of graph properties [21]; and (v) long-term network properties, e.g., price equilibrium, pandemic thresholds [16]. In addition, there has been a lot of research in studying networks of information resources, in particular looking at search and navigation in networks [18, 33, 1]. Other research problems in this area include studying and predicting how clusters and referral chains of people with related interests form in social networks [17], the role of geographical proximity in social networks in facilitating friendship [22], modeling trust and reputation in networks [15], link analysis to find relative importance of network nodes [11], and learning models for topic, role, and group patterns in social networks [23].

In related work, there has been recent research on modeling the stochastic co-evolution of structure and strategy in networks using a reinforcement learning model in a repeated game setting [26]. Techniques from optimal control and decision theory have been used to empirically study dynamics of learning time-varying network parameters [6]. There have been studies of the characteristics of social dynamics behind peer production systems like collaborative content generation (e.g., Wikipedia) and open source software design (e.g., Linux), to understand the circumstances under which such collaborative social endeavors become successful [12, 8]. Query routing [30] and search [31] have also been studied in peer-to-peer networks.

The novelty of our approach is applying online learning techniques for social search in a dynamic message routing system, with rich node and link models and a centralized supernode. The learning framework can find different interesting emerging network properties, e.g., find sub-groups of networked individuals who are most effective at content generation, suggest links between people who have related interests profiles, detect and track emerging communities of interest, and so on. We also propose a way not only to analyze peer production behavior but to facilitate formation of new connections between nodes in a dynamic network, using suitable online machine learning and data mining tools.

On the application side, there have been some recent tools that perform Q/A over a social network, notably Yahoo! Answers and Yedda. However, iLink is more general than these applications, since they do not allow routing of questions within the social network (their model is similar to discussion boards or forums), and cannot handle general messages (only support questions and answers).

9. FUTURE RESEARCH

Each of the algorithmic solutions provided for the issues listed in Section 5 can be possibly improved by better approaches. Apart from those improvements, the model formulation for iLink opens up several other interesting areas of future research related to content search and message routing in social networks:

1. **Hierarchical topics**: The topic model implementation should be generalized to handle hierarchies, or better still an arbitrary DAG structure over the topics. In that case, hierarchical prior knowledge, e.g., ontologies, organization charts, can be directly incorporated as hierarchical priors into the topic model.

2. **Online and temporal topic inference**: In some real cases, the total number of relevant topics in the topic set may not be known a priori, because of data sparsity and lack of domain knowledge. In such situations, it is not sufficient to only update the existing topic distributions with new messages or new users, as done currently in Section 5 – it would be preferable to have a model selection algorithm that can decide when not to update the existing topics but rather add a new topic. We want to design and implement a model selection technique that would incrementally take into account how topics change over time, combining temporal topic learning algorithms with model selection in the online learning setting.

3. **Distributed validation**: In the FAQtory system, the node asking the question may not know the correctness of a suggested answer, or multiple answers may need to be compared in terms of quality. In such situations, a distributed voting scheme is used for rating the questions, where the nodes who participated in the message routing can each vote on the validity of an answer. The interesting research problem here is to get a consensus on the best answer from distributed votes while factoring in the intrinsic reliability of the vote cast by a node, and suitably updating the voting reliability score for a node after the consensus has been reached.

4. **Noise robustness**: Any algorithm operating online within a large-scale practical social network learning system like iLink needs a mechanism for being robust to possible sources of noise. In the future, we want to implement spam filtering techniques, especially for controlling content spam in messages and link spam in message routing.

5. **Rank function learning**: The overall rank score (Section 3) is a weighted linear combination of the different profile components. Currently, the weights on each parameter are considered as inputs to the ranking algorithm. However, the weights of such a score function can be learned in an online fashion from voting feedback, using a multiplicative update algorithm to combine scores from multiple experts.

6. **Incentive model**: In the present instantiation of iLink, we have employed a reputation economy model based on adhoc expert group members rating messages that are submitted for review. In the future, we intend to incorporate more detailed incentive models, similar to recent work in peer-to-peer incentive models [10, 19]. An interesting research problem is to design incentive models and an economy of social capital where there are heterogeneous distributed human nodes as well as a global system supernode.

7. **Dynamic social network modeling**: One promising area of related research deals with learning and exploiting key characteristics as social networks change in time. Our proposed online learning framework can be extended to sense and act upon emerging signals, for instance correcting rank weights and message routing paths for particular circumstances. Finding scalable methods for detecting emerging network changes, such as clusterability, cohesion, centrality, prestige, or triadic structural shifts, could yield rich inputs to the larger iLink system.

10. CONCLUSIONS

The web has made unprecedented peer production frameworks possible. These frameworks are unprecedented in terms of both scale and intensity. If this model is going to be successfully extended to a wider range of activities and goals, we believe that it will be essential to understand how the underlying social networks operate. To accomplish this, we have first introduced a model of social networks that incorporates important node and link information. We have then proposed a specific model iLink of peer production by message passing, and discussed the social generation of a FAQ repository using a system FAQtory that instantiates this model. By introducing a supernode that learns how the social network solves the message routing problem, we are able to learn patterns of social interaction in this message routing model. The FAQtory application supports both the generation of QA pairs and the follow-up validation/modification (analogous to the Wikipedia edit) capabilities. This basic approach has a number of extensions to other applications and peer production models, including RSS and general social search. We believe that our approach, which focuses on the underlying message passing dynamics, is a promising avenue for the important task of extending the peer production paradigm. Clearly, many practical and theoretical issues remain. The importance of making such models work for real, ongoing web-based collaboration efforts (e.g., the military knowledge sharing portals) is a growing issue and a wonderful challenge for machine learning and data mining techniques.

Acknowledgment: The authors would like to thank Mark Drummond, Arindam Banerjee and Colin Evans for valuable feedback. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA, or the Department of Interior-National Business Center (DOI-NBC).

11. REFERENCES

[1] L. A. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187–203, July 2005.

[2] A. Banerjee and S. Basu. A social query model for decentralized search. Technical report, University of Minnesota, 2007.

[3] A. Banerjee and S. Basu. Topic models over text streams: A study of batch and online unsupervised learning. In *Proc. of the 7th SIAM Intl. Conf. on Data Mining*, 2007.

[4] A. Banerjee, S. Basu, and S. Merugu. Multiway clustering on relation graphs. In *Proc. of the 7th SIAM Intl. Conf. on Data Mining*, 2007.

[5] S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In *Proc. of 10th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD-2004)*, pages 59–68, 2004.

[6] G. W. Beck and V. Wieland. Learning and control in a changing economic environment. *Journal of Economic Dynamics and Control*, 26(9-10):1359–1377, August 2002.

[7] R. Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics 6*, pages 679–684, 1957.

[8] Y. Benkler. Coase's penguin, or Linux and the nature of the firm. *Yale Law Journal*, 112, 2002.

[9] Y. Benkler. *The Wealth of Networks.* Yale University Press, 2006.

[10] L. Y.-K. Blanc, A. and A. Vahdat. Designing incentives for peer-to-peer routing. In *2nd wokshop on Economics of Peer-to-peer Systems*, 2004.

[11] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

[12] R. Cross, A. Parker, L. Prusak, and S. Borgatti. Knowing what we know: Supporting knowledge creation and sharing in social networks. *Organizational Dynamics*, 302(2):100–120, 2001.

[13] I. Dhillon, S. Mallela, and R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3(4):1265–1287, 2003.

[14] R. A. Ghosh. Cooking pot markets: An economic model for the trade in free goods and services on the internet. *First Monday*, 3(3), 1998.

[15] T. Hogg and L. A. Adamic. Enhancing reputation mechanisms via online social networks. In *Proc. 5th ACM conference on Electronic Commerce*, pages 236–237, 2004.

[16] S. M. Kakade, M. Kearns, L. E. Ortiz, R. Pemantle, and S. Suri. Economic properties of social networks. In *Advances in Neural Info. Processing Systems 17*, 2005.

[17] H. Kautz, B. Selman, and M. Shah. Referral web: Combining social networks and collaborative filtering. *Commun. ACM*, 40(3):63–65, 1997.

[18] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proc. 32nd ACM Symposium on Theory of Computing*, 2000.

[19] J. Kleinberg and P. Raghavan. Query incentive networks. In *Proc. 46th Annual IEEE Symposium on Foundations of Computer Science*, 2005.

[20] J. Lerner and J. Tirole. Some simple economics of open source. *Journal of Industrial Economics*, 50(2):197–234, 2002.

[21] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery in Data Mining.*

[22] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proc. National Academy of Sciences*, 102(33), 2005.

[23] A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and role discovery in social networks. In *Proc. of Intl. Joint Conf. on AI*, 2005.

[24] S. Merugu and J. Ghosh. A privacy-sensitive approach to distributed clustering. *Pattern Recognition Letters*, 26:399–410, 2005.

[25] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[26] R. Pemantle and B. Skyrms. A dynamic model of social network formation. *Proc. of the National Academy of Sciences*, 2000.

[27] E. Raymond. *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary.* O'Reilly & Associates, Sebastopol, CA, 1999.

[28] P. Sarkar and A. Moore. Dynamic social network analysis using latent space models. *SIGKDD Explorations: Special Edition on Link Mining*, 2005.

[29] D. Tapscott and A. Williams. *Wikinomics: How mass collaboration changes everything.* Portfolio, 2006.

[30] C. Tempich, S. Staab, and A. Wranik. Remindin': Semantic query routing in peer-to-peer networks based on social metaphors. In *Proc. of Intl. WWW Conf.*, 2004.

[31] D. Tsoumakos and N. Roussopoulos. Adaptive probabilistic search for peer-to-peer networks. In *3rd IEEE Intl. Conf. on P2P Computing*, 2003.

[32] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[33] D. J. Watts, P. S. Dodds, and M. E. J. Newman. Identity and search in social networks. *Science*, 296:1302–1305, 2002.

[34] F. Wu, B. A. Huberman, L. A. Adamic, and J. R. Tyler. Information flow in social groups. *Physica A*, 337:327–335, 2004.