Leveraging Social Foci for Information Seeking in Social Media

Suhas Ranganath
srangan8@asu.edu
Arizona State University

Jiliang Tang
Jiliang.Tang@asu.edu
Arizona State University

Xia Hu
Xia.Hu@asu.edu
Arizona State University

Hari Sundaram
hs1@illinois.edu
University of Illinois Urbana Champaign

Huan Liu
Huan.Liu@asu.edu
Arizona State University

Abstract

The rise of social media provides a great opportunity for people to reach out to their social connections to satisfy their information needs. However, generic social media platforms are not explicitly designed to assist information seeking of users. In this paper, we propose a novel framework to identify the social connections of a user able to satisfy his information needs. The information need of a social media user is subjective and personal, and we investigate the utility of his social context to identify people able to satisfy it. We present questions users post on Twitter as instances of information seeking activities in social media. We infer soft community memberships of the asker and his social connections by integrating network and content information. Drawing concepts from the social foci theory, we identify answerers who share communities with the asker w.r.t. the question. Our experiments demonstrate that the framework is effective in identifying answerers to social media questions.

Information seeking is defined as "A conscious effort to acquire information in response to a need or gap in knowledge" (Case 2012). Online social media makes it easier for users to reach out to a large number of friends, leading people to use them to seek information from their social connections. This gives rise to a distinct way for online information seeking, wherein the information needs expressed are subjective and personal to the asker. An interesting way people leverage online social media to seek information is by asking questions through their status messages (Morris, Teevan, and Panovich 2010). This phenomenon is prevalent in social media platforms like Twitter and Facebook and has received considerable attention in recent literature (Efron and Winget 2010a; Paul, Hong, and Chi 2011; Lampe et al. 2014).

However, unlike dedicated Q&A platforms, generic social media sites like Twitter and Facebook are not designed for information seeking (Paul, Hong, and Chi 2011). Questions are not archived, thus finding people who answered similar questions in the past is difficult. Questions are buried among other content produced by the social connections of a potential answerer. Designing algorithmic frameworks to identify answerers to social media questions will help to bridge the information gap of users and increase user satisfaction. This framework can also help enhance Twitter search by making it personalized to the asker.

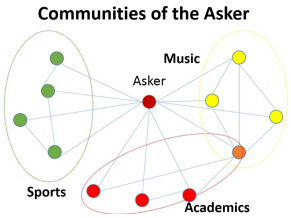
Information need of social media user is subjective or personal, unlike traditional Q&A platforms like Stackoverflow, and hi social context is useful to find appropriate people able to satisfy it (Hecht et al. 2012). Also, users with higher tie strength with the asker were shown to better satisfy information needs in social media (Panovich, Miller, and Karger 2012). For example, to assist a person looking to get a new hairstyle, finding people from his social connections who share related context with him can be more useful to him than finding web pages related to hair salons.

This task faces several challenges. The questions are textual while the social context of the asker can involve network information. Integrating such kind of heterogeneous information will help to efficiently utilize social context to identify answerers to social media questions. Each social media user has many social connections and produces a lot of content leading to significant issues of scalability. Finally, the social context of the asker related to the question needs to be determined and appropriately utilized in order to identify suitable answerers.

In sociological literature, the social foci theory postulates that interactions between people are organized around relevant entities known as foci (Feld 1981). A focus can be the activities, interests, and various affiliations of a user. Different groups of social connections of a user share different foci with him. For example, from Fig. 1(a) we see that the user shares an interest of sports with his connections in green, an interest of music with his connections in yellow and academic interests with his connections in red.

Inspired by the social foci theory we propose that, people in social media sharing social foci related to the question with the asker are suitable to answer them. Illustrative examples of questions are given in Fig. 1(b). The asker of Q1 is seeking assistance in his math homework, and this might be best responded by users sharing academic foci with him. Q2 is seeking opinions on an NFL game, and this might be best provided by his connections sharing foci related to sports with the asker. Similarly, Q3 might be best answered by connections sharing music related foci with the asker.

In this paper, we propose a framework to investigate the utility of social context derived from network and content



(a)

Questions
Q1:Who can help me with my math homework?
Q2: Thoughts on the next NFL game? I'm undecided
Q3: Need to update my playlist. Any suggestions?

(b)

Figure 1: (a) Different foci a user shares with his social connections. (b) Questions of users. Users sharing different foci with the asker are more likely to answer related questions.

information in identifying answerers to social media questions. Informed by the concepts of social foci theory illustrated in Fig. 1, we utilize social context of an asker related to the question, and demonstrate that the framework is effective in identifying answerers for social media questions. Specifically, we address the following questions: How to utilize the network and content information of the asker and his social connections to better identify answerers for social media questions? Are approaches based on shared context in the question domain useful in identifying answerers to different kinds of social media questions?

The main contributions of our work are as follows:

- Formally defining the problem of finding suitable users to answer questions in online social media platforms,
- Proposing a framework to exploit network and content information to identify answerers to questions, and
- Conducting experimental evaluations of the framework on a dataset of social media questions.

Related Work

Social media questions have received considerable attention in research communities (Yang et al. 2011; Efron and Winget 2010b; Lee et al. 2012). An analytical study on questions asked and the answers received in Twitter is presented in (Morris, Teevan, and Panovich 2010; Paul, Hong, and Chi 2011). They indicated that subjective questions were the most prevalent and the trust users have on their friends was the primary factor for asking questions. A study of questions and responses received in Facebook was conducted in (Gray et al. 2013; Ellison et al. 2013) and bridging social capital was proposed to be a strong motivation for Q&A activity in social media. These works give interesting insights to the question answering process in social media, but do not focus on identifying answerers to these questions.

Systems to identify answerers for social media questions adopt different methods such as matching question content with profile information (Hecht et al. 2012) and using crowdsourced technology (Jeong et al. 2013). Social search architectures and empirical models to route questions to answerers using different kinds of social information are discussed in (Horowitz and Kamvar 2010; Nandi et al. 2013). These works are meant to demonstrate social search systems and hence do not contain any experimental evaluations.

A related line of research is the study of community Q&A systems like Yahoo! Answers (Adamic et al. 2008) and Quora (Wang et al. 2013a). Content from existing Q&A sessions are used to rank answerers by NLP techniques. (Jurczyk and Agichtein 2007) uses link structure to find authoritative answerers for a question category. (Zhou et al. 2012) and (Yang et al. 2013) combine network and content information to identify authoritative users as answerers. The environment for social media questions is different as the candidate answerers are themselves connected via social relations. Systems utilizing question categories (Zhu et al. 2013) cannot be applied as they are not explicitly known in generic social media. Social expertise systems (Pal and Counts 2011; Bozzon et al. 2013) identify subject matter experts in social media. Social media questions are subjective and personal might require answerers who share social context with the asker rather than subject matter experts.

Another related field to our work is the application of social foci theory in social media. Social foci theory has received attention in several domains such as relational learning (Tang and Liu 2009) and structural hole theory (Burt 2009). Recently, social foci theory has been used to derive community memberships using both node and edge attributes (Yang, McAuley, and Leskovec 2013). To the best of our knowledge, this is the first work that has utilized concepts from social foci theory to identify answerers for social media questions.

Problem Definition

We first describe some general notations. Boldface uppercase letters (e.g $\mathbf{X}$) denote matrices and boldface lowercase letters (e.g. $\mathbf{x}$) denote vectors. $\mathbf{X}_{ij}$ signifies the element in the i$^{\text{th}}$ row and j$^{\text{th}}$ column of matrix $\mathbf{X}$ and the i$^{\text{th}}$ row of a matrix $\mathbf{X}$ is denoted by $\mathbf{X}_i$. Similarly $\mathbf{x}_i$ denotes the i$^{\text{th}}$ element of vector $\mathbf{x}$ and $\mathbf{x}_y$ denotes a vector $\mathbf{x}$ corresponding to the quantity y. We denote the Frebonius norm of a matrix $\mathbf{X}$ as $||\mathbf{X}||_F = \sqrt{\sum_{i,j} \mathbf{X}_{ij}^2}$.

We now define some terms related to the questions asked, the network and content of the asker and his social connections. We define attributes of a question q as the set of words used in the question i.e. $\mathbf{w}_q = [w_{q1}, w_{q2}, ..., w_{ql}]$. Since we are dealing with subjective questions, the asker marking the answer to be useful or publicly acknowledging the answerer gives the evidence of its acceptance.

Let A denote the asker of the question q and $\mathbf{f}_A = [f_1, f_2, ..., f_m]$ denote the social connections of A and m is the number of social connections of A. We define the egonetwork of each asker A as consisting of the asker, the social connections of the asker and the links among his social connections. The egonetwork of asker A, $\mathbf{N} \in \mathbb{R}^{(m+1)\times(m+1)}$ is given by

$$\mathbf{N}_{ij} = \begin{cases} 1 & \text{directed edge from } f_j \text{ to } f_i, i \neq j, i,j \in \{A, f_A\} \\ 0 & \text{otherwise} \end{cases}$$

We collect the status messages of the asker and his social connections. We apply basic preprocessing steps such as removal of stop words and stemming. We then define the user-word matrix $\mathbf{S} \in \mathbb{R}^{(m+1)\times w}$ of asker A as

$$\mathbf{S}_{ij} = \begin{cases} \text{num*tfidf}_j & \text{if user } u_i \text{ has used word } w_j \text{ num times} \\ 0 & \text{if user } u_i \text{ has not used the word } w_j, \end{cases}$$

where num is the number of times the user u_i has used the word w_j, w is the total number of words used by the asker and his social connections and $tfidf_j$ is the tf-idf score of word w_j. A single user will only use a small subset of the total number of words, resulting in $\mathbf{S}$ being sparse.

With the terminologies and the notations described above, we formally define the problem as follows *"Given a question q, an asker A, the network neighborhood of the asker $\mathbf{f}_A$, find a suitable set of people among $\mathbf{f}_A$ whose responses for the question q that the asker accepts"*.

Information Seeking via Social Foci

In this section, we describe our framework to identify answerers for social media questions in detail. First, we infer social foci memberships of the asker and his social connections from their network and content information. We then compute the overlap in foci memberships of the asker and his social connections in the question domain to identify answerers to these questions.

Modeling Content Information

We model the content information to infer major foci of the asker and his social connections. We draw from Non-negative Matrix Factorization (NMF) presented in (Seung and Lee 2001) to infer foci from the user-word matrix $\mathbf{S} \in \mathbb{R}^{(m+1)\times w}$. We factorize the matrix $\mathbf{S}$ into two low dimensional sparse non-negative matrices, $\mathbf{U} \in \mathbb{R}^{(m+1)\times k}$ and $\mathbf{P} \in \mathbb{R}^{w\times k}$ such that k $\ll$ m by solving the following optimization problem.

$$\min_{\mathbf{U}\geq 0,\mathbf{P}\geq 0} ||\mathbf{S} - \mathbf{U}\mathbf{P}^{\mathrm{T}}||_F^2 \tag{1}$$

Here, k is the number of latent foci in the neighborhood of the asker and m is the number of his social connections. $\mathbf{U}$ denotes the latent foci membership of the asker and his social connections and $\mathbf{P}$ denotes the latent foci memberships of words. The correlation between foci memberships of the words can be obtained by the overlap in the corresponding rows of $\mathbf{P}$. The constraints $\mathbf{U} \geq 0$ and $\mathbf{P} \geq 0$ denote that the matrices have all non-negative elements. The non-negativity ensures an intuitive decomposition of the matrix into its constituent parts.

Integrating Network Information

In a social setting, the interests or affiliations of an user are correlated with the interests of his social connections, thereby affecting his memberships to different foci (Feld 1981). This notion is also supported by network homogeneity (Marsden 1988), which says that people connected to each other display similar interests and affiliations. Therefore, it is essential to utilize network structure to determine foci memberships of the asker and his social connections.

To utilize the network structure, we first factorize the ego network of the asker $\mathbf{N}$ into two low rank non-negative matrices $\mathbf{U} \in \mathbb{R}^{(m+1)\times k}$ and $\mathbf{V} \in \mathbb{R}^{k\times k}$ s.t. k $\ll$ m by solving the following optimization problem.

$$\min_{\mathbf{U}\geq 0,\mathbf{V}\geq 0} ||\mathbf{N} - \mathbf{U}\mathbf{V}\mathbf{U}^{\mathrm{T}}||_F^2, \tag{2}$$

where $\mathbf{U}$ contains the membership of the asker and his social connections to different latent foci and $\mathbf{V}$ contains the correlations between the foci. The constraints $\mathbf{U} \geq 0$ and $\mathbf{V} \geq 0$ denote that the matrices have only non-negative elements.

We then integrate network and content information to infer the foci membership of the asker and his social connections by formulating the following optimization problem.

$$\begin{aligned}\min_{\mathbf{U}\geq 0,\mathbf{V}\geq 0,\mathbf{P}\geq 0} \quad & \alpha||\mathbf{S} - \mathbf{U}\mathbf{P}^{\mathrm{T}}||_F^2 + \beta||\mathbf{N} - \mathbf{U}\mathbf{V}\mathbf{U}^{\mathrm{T}}||_F^2 \\ & +\gamma(||\mathbf{U}||_F^2 + ||\mathbf{V}||_F^2 + ||\mathbf{P}||_F^2)\end{aligned} \tag{3}$$

Here $\mathbf{U}$ contains the latent foci membership of the asker and his connections obtained by integrating network and content information, $\mathbf{P}$ shows the latent foci memberships of the words and $\mathbf{V}$ represents the correlation between the latent foci. $||\mathbf{U}||_F^2$, $||\mathbf{V}||_F^2$, and $||\mathbf{P}||_F^2$ are regularization terms introduced to prevent overfitting and γ is the positive parameter for control the proportions of the regularization terms. The constraints $\mathbf{U} \geq 0$, $\mathbf{V} \geq 0$, and $\mathbf{P} \geq 0$ denote that the matrices do not contain negative elements. α and β are positive parameters to control the effects of content and network proportions respectively.

We draw from the concepts of the social foci theory illustrated in Fig. 1 to propose that users sharing a large amount of foci memberships with the asker in the question domain can effectively answer social media questions. The shared foci memberships of the asker with his social connections is given by the overlap between their corresponding rows in $\mathbf{U}$. The question domain in the latent foci space is obtained by combining the rows of $\mathbf{P}$ corresponding to the words in the question. Before formalizing these notions, we optimally derive the latent matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{P}$ by solving Eq. (3).

Deriving the Optimal Latent Matrices

The problem presented in Eq. (3) belongs to a class of constrained convex minimization problems. Motivated by (Ding et al. 2006), we describe an algorithm to find optimal solutions for $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{P}$. The key idea is to optimize the objective with respect to one variable while fixing others. The three variables are iteratively updated until convergence.

From Eq.(3), we let

$$\begin{aligned}\mathcal{J} = \; & \alpha||\mathbf{S} - \mathbf{U}\mathbf{P}^{\mathrm{T}}||_F^2 + \beta||\mathbf{N} - \mathbf{U}\mathbf{V}\mathbf{U}^{\mathrm{T}}||_F^2 + \\ & \gamma(||\mathbf{U}||_F^2 + ||\mathbf{V}||_F^2 + ||\mathbf{P}||_F^2)\end{aligned} \tag{4}$$

We then take the Lagrangian of the objective function $\mathcal{J}$. Let the Lagrange multiplier for the constraints $\mathbf{U} \geq 0$, $\mathbf{V} \geq 0$, and $\mathbf{P} \geq 0$ be Λ_u, Λ_v, and Λ_p respectively. Then

$$\mathcal{L} = \mathcal{J} + tr(\Lambda_u\mathbf{U}^{\mathrm{T}}) + tr(\Lambda_v\mathbf{V}^{\mathrm{T}}) + tr(\Lambda_p\mathbf{P}^{\mathrm{T}}) \tag{5}$$

We compute the partial derivatives of the lagrangian $\mathcal{L}$ with respect to $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{P}$ keeping the other variables fixed as shown below.

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{U}} = \; & 2(\alpha(-\mathbf{S}\mathbf{P} + \mathbf{U}\mathbf{P}^{\mathrm{T}}\mathbf{P}) + \beta(-\mathbf{N}^{\mathrm{T}}\mathbf{U}\mathbf{V} - \mathbf{N}\mathbf{U}\mathbf{V}^{\mathrm{T}} \\ & + \mathbf{U}\mathbf{V}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{V}^{\mathrm{T}} + \mathbf{U}\mathbf{V}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{V}) + \gamma\mathbf{U}) + \Lambda_u \\ \frac{\partial\mathcal{L}}{\partial\mathbf{V}} = \; & 2(\beta(-\mathbf{U}^{\mathrm{T}}\mathbf{N}\mathbf{U} + \mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{V}\mathbf{U}^{\mathrm{T}}\mathbf{U}) + \gamma\mathbf{V}) + \Lambda_v \\ \frac{\partial\mathcal{L}}{\partial\mathbf{P}} = \; & 2(\alpha(-\mathbf{S}^{\mathrm{T}}\mathbf{U} + \mathbf{P}\mathbf{U}^{\mathrm{T}}\mathbf{U}) + \gamma\mathbf{P}) + \Lambda_p.\end{aligned} \tag{6}$$

Substituting the KKT complementary conditions in Eq. (6) and rearranging we get the following update rules for latent matrices $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{P}$.

$$\mathbf{U}_{ij} \leftarrow \mathbf{U}_{ij}\sqrt{\frac{\alpha\mathbf{SP}+\beta(\mathbf{N}^{\mathrm{T}}\mathbf{UV}+\mathbf{NUV}^{\mathrm{T}})}{\alpha\mathbf{UP}^{\mathrm{T}}\mathbf{P}+\beta(\mathbf{UVU}^{\mathrm{T}}\mathbf{UV}^{\mathrm{T}}+\mathbf{UV}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{UV})+\gamma\mathbf{U}}}$$

$$\mathbf{V}_{ij} \leftarrow \mathbf{V}_{ij}\sqrt{\frac{\beta\mathbf{U}^{\mathrm{T}}\mathbf{NU}}{\beta(\mathbf{U}^{\mathrm{T}}\mathbf{UVU}^{\mathrm{T}}\mathbf{U})+\gamma\mathbf{V}}}$$

$$\mathbf{P}_{ij} \leftarrow \mathbf{P}_{ij}\sqrt{\frac{\alpha\mathbf{S}^{\mathrm{T}}\mathbf{U}}{\alpha\mathbf{PU}^{\mathrm{T}}\mathbf{U}+\gamma\mathbf{P}.}} \tag{7}$$

The optimization algorithm is summarized in Steps 1-7 in **Algorithm 1**. The square root on the update rules is added to ensure convergence (Ding, Li, and Jordan 2008). The correctness and convergence of the rules can be proved by the axillary function method (Lee and Seung 2000).

Identifying Answerers from Foci Information

We now identify relevant answerers from the social connections of the asker using the latent matrices $\mathbf{U}$,$\mathbf{V}$ and $\mathbf{P}$. We first extract the words from the question attribute vector $\mathbf{w}_{\mathrm{q}}$ and obtain the foci memberships of each word from the corresponding rows in matrix $\mathbf{P}$. We then compute the domain of the question in the latent foci space as a combination of individual word membership vectors as

$$\mathbf{d}_{\mathrm{q}} = \sum_{\mathrm{w_i}\epsilon\mathbf{w}_{\mathrm{q}}} \mathbf{P}_i, \tag{8}$$

where $\mathbf{d}_{\mathrm{q}}$ represents the domain of the question q in the latent foci space and $\mathrm{w_i}$ is the word corresponding to the i^{th} row of $\mathbf{P}$.

We next compute the foci memberships of the asker and his social connections in the question domain. The Hadamard product of two vectors is the pointwise product of their respective elements, and it exactly captures this notion. For each question, we compute the Hadamard product of the row of $\mathbf{U}$ corresponding to the asker, $\mathbf{U}_{\mathrm{A}}$ and the vector representing the question domain $\mathbf{d}_{\mathrm{q}}$.

$$\mathbf{g}_{\mathrm{A}} = \mathbf{U}_{\mathrm{A}} \circ \mathbf{d}_{\mathrm{q}}, \tag{9}$$

where $\mathbf{g}_{\mathrm{A}}$ contains the foci membership of the asker in the domain of the question. Similarly, we compute the foci memberships of each social connection of the asker in the domain of the question q by

$$\mathbf{g}_{\mathrm{f_m}} = \mathbf{U}_{\mathrm{f_m}} \circ \mathbf{d}_{\mathrm{q}}, \tag{10}$$

where $\mathrm{f_m}$ is the m^{th} social connection of the asker, $\mathbf{U}_{\mathrm{f_m}}$ is the row of matrix $\mathbf{U}$ corresponding to $\mathrm{f_m}$ and $\mathbf{g}_{\mathrm{f_m}}$ contains the foci membership of $\mathrm{f_m}$ w.r.t the domain of the question.

Finally, we find the overlap in foci memberships of the asker and his social connections in the question domain as

$$\mathbf{rs}(\mathrm{q},\mathrm{A},\mathrm{f_m}) = sim(\mathbf{g}_{\mathrm{A}}, \mathbf{g}_{\mathrm{f_m}}), \tag{11}$$

where $\mathbf{rs}(\mathrm{q},\mathrm{A},\mathrm{f_m})$ denotes the score of the answerer $\mathrm{f_m}$ to the question q by the asker A. We sort the answerers according to their score and return them to the asker as a ranked list, $\mathbf{ra}$. Results with different similarity metrics is presented in Table 2. The method for identifying answerers from foci information is summarized in Steps 8-11 in **Algorithm 1**. The quantity $\mathbf{rs}(\mathrm{q},\mathrm{A},\mathrm{f_m})$ signifies the context in terms of network and content shared between asker A and his social connection $\mathrm{f_m}$ in the domain of question q.

Algorithm 1: Automatic Identification of Answerers to Social Media Questions

Input: Question q of asker (A), friends and followers of A, $\mathbf{f}_{\mathrm{A}} = [\mathrm{f}_1, \mathrm{f}_2, ..., \mathrm{f_m}]$, Egonetwork of the asker (**N**), user-word matrix of the asker and his connections (**S**) and $\{\alpha, \beta, \gamma, k\}$
Output : A ranked list of the potential answerers $\mathbf{ra}$
1: Initialize $\mathbf{U}$, $\mathbf{V}$, $\mathbf{P}$ randomly
2: **while** not convergent **do**
3: update
4: $\mathbf{U}_{ij} \leftarrow \mathbf{U}_{ij}\sqrt{\frac{\alpha\mathbf{SP}+\beta(\mathbf{N}^{\mathrm{T}}\mathbf{UV}+\mathbf{NUV}^{\mathrm{T}})}{\alpha\mathbf{UP}^{\mathrm{T}}\mathbf{P}+\beta(\mathbf{UVU}^{\mathrm{T}}\mathbf{UV}^{\mathrm{T}}+\mathbf{UV}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{UV})+\gamma\mathbf{U}}}$
5: $\mathbf{V}_{ij} \leftarrow \mathbf{V}_{ij}\sqrt{\frac{\beta\mathbf{U}^{\mathrm{T}}\mathbf{NU}}{\beta(\mathbf{U}^{\mathrm{T}}\mathbf{UVU}^{\mathrm{T}}\mathbf{U})+\gamma\mathbf{V}}}$
6: $\mathbf{P}_{ij} \leftarrow \mathbf{P}_{ij}\sqrt{\frac{\alpha\mathbf{S}^{\mathrm{T}}\mathbf{U}}{\alpha\mathbf{PU}^{\mathrm{T}}\mathbf{U}+\gamma\mathbf{P}}}$
7: **end while**
8: $\mathbf{w}_{\mathrm{q}} = [\mathrm{w_{q1}}, \mathrm{w_{q2}}, ..., \mathrm{w_{ql}}]$, $\mathbf{d}_{\mathrm{q}} = \sum_{\mathrm{w_i}\epsilon\mathbf{w}_{\mathrm{q}}} \mathbf{P}_i$
9: $\mathbf{g}_{\mathrm{A}} = \mathbf{U}_{\mathrm{A}} \circ \mathbf{d}_{\mathrm{q}}$, $\mathbf{g}_{\mathrm{f_m}} = \mathbf{U}_{\mathrm{f_m}} \circ \mathbf{d}_{\mathrm{q}}$
10: $\mathbf{rs}(\mathrm{q},\mathrm{A},\mathrm{f_m}) = sim(\mathbf{g}_{\mathrm{A}}, \mathbf{g}_{\mathrm{f_m}})$
11: $\mathbf{ra}$=sort($\mathbf{rs}$)

Time Complexity

The highest time cost results from updating the latent matrices in steps 4-6. In the updating terms, the complexity of the terms $\mathbf{SP}$ and $\mathbf{S}^{\mathrm{T}}\mathbf{U}$ is low due to sparsity of $\mathbf{S}$. The terms $\mathbf{N}^{\mathrm{T}}\mathbf{UV}$, $\mathbf{NUV}^{\mathrm{T}}$ and $\mathbf{U}^{\mathrm{T}}\mathbf{NU}$ have a complexity of $\mathrm{O}(\mathrm{mk}^2)$ where m is the number of friends and k is the number of latent dimensions due to the sparsity of $\mathbf{N}$. The terms $(\mathbf{U}(\mathbf{V}(\mathbf{U}^{\mathrm{T}}\mathbf{U})\mathbf{V}^{\mathrm{T}})$, $(\mathbf{U}(\mathbf{V}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}}\mathbf{U})\mathbf{V})$ and $((\mathbf{U}^{\mathrm{T}}\mathbf{U})\mathbf{V}(\mathbf{U}^{\mathrm{T}}\mathbf{U}))$ has a complexity of $\mathrm{O}(\mathrm{mk}^2)$ when computed as shown in the brackets. The complexity of $\mathbf{PU}^{\mathrm{T}}\mathbf{U}$ and $\mathbf{UP}^{\mathrm{T}}\mathbf{P}$ is $\mathrm{O}((\mathrm{w+m})\mathrm{k}^2)$ where w is the number of words. Therefore, the overall complexity of a single iteration is $\mathrm{O}((\mathrm{w+m})\mathrm{k}^2)$, which is low owing to the few number of latent dimensions. In addition, notice that steps 1-7 can be computed offline and only steps 8-10 are computed when the question is asked, further reducing the time required to identify answerers for a given question.

Experiments

In this section, we first present a dataset of questions posted on Twitter and then conduct experiments to answer the following questions that help in understanding the framework better: How does the proposed framework perform in comparison to existing baselines? What is the effect of the amount of network and content information on the performance of the framework?

Dataset

The dataset consists of subjective questions from the social media platform Twitter. We follow the literature on questions in Twitter (Morris, Teevan, and Panovich 2010) to construct a keyword set related to subjective questions. We append "?" to each keyword to collect questions from the Twitter Streaming API. Texts having "?" in online content

Parameter	Statistics
# of Questions	1065
# of Askers	1026
# of Selected Answers	1450
# of Followers and Friends of the askers	966,117
Median # of Followers and Friends per asker	588
Median # Tweets per user	479

Table 1: Dataset containing questions posted in Twitter with statistics related to network and content information.

are shown to be questions with high precision (Cong et al. 2008). We deem replies to have been accepted by the asker if he has marked it as "favorite" or acknowledged the answerer by using "thanks" or "thank you". We mark the users who provided these answers as the ground truth for each question following (Hecht et al. 2012). Some important statistics of the dataset are given in Table 1. The first question was posted on Dec 27, 2013 and the last one on Jan 15, 2014. We use the methods in the public Twitter API to collect the friends, followers and public status messages of the asker to obtain the asker's social connections and their interests (Kumar, Morstatter, and Liu 2013). We use the data to construct the ego network $\mathbf{N}$ and user-word matrix $\mathbf{S}$ for each asker.

Experiment Settings

We introduce the following metrics to evaluate the performance of our framework: The Mean Reciprocal Rank (MRR) (Radev et al. 2002) is a measure of the overall likelihood of the framework to identify an answerer for a question, the Mean Average of Precision (MAP) (Bian et al. 2008) measures the potential satisfaction of the asker with the top K results and the Normalised Discounted Cumulative Gain (NDCG)@K considers the order within the top K rankings (Wang et al. 2013b).We use the following baselines to evaluate the performance of our framework.

Random: We randomly order the friends and followers of the asker 100 times and return the mean ordering.

Aardvark (Horowitz and Kamvar 2010): This paper describes a search engine which directed questions posted on the system to users with formulation to compute affinity with the asker and interest in the question topics. It does not consider the network structure and also does not contain experimental evaluations of its formulation.

Content based Methods (Riahi et al. 2012): The paper focuses on community Q&A like Yahoo! Answers and compares the similarity of the question topic with the interests of the answerers derived only from the their content. The interests were inferred by two topic models: LDA and the Segmented Topic Model (STM) (Du, Buntine, and Jin 2010).

Topic Sensitive Page Rank (Zhou et al. 2012): This paper employs a PageRank based approach to find subject matter experts in the question topic by combining network and content information of the potential answerers. The paper identifies topical authorities not considering the shared context between the asker and the answerers.

Shared Foci: This baseline measures the effect of shared user context. It computes the shared foci memberships of the asker and his social connections derived from either network (α=0) or content (β=0) information. The question information is not taken into consideration. This also helps in evaluating methods using only network structure.

Method	MRR	MAP@5	NDCG@5
Random	1.20%	1.12%	0.25%
Content-LDA	1.56%	1.46%	0.30%
Content-STM	1.93%	2.27%	0.50%
TSPR	1.64%	1.63%	0.45%
Aardvark	2.11%	2.53%	0.50%
Shared Foci (Network)	3.43%	3.66%	0.97%
Shared Foci (Content)	3.60%	3.87%	1.17%
Our Model (Cosine)	3.91%	4.63%	1.25%
Our Model (PCC)	3.80%	4.73%	1.31%
Our Model (Euclidean)	4.36%	5.54%	1.41%

Table 2: Comparison of performance of the proposed framework with baselines.

For initial experiments, we set the parameters in Eq. (3) as follows. The regularization parameter is set at γ=0.01. The number of topics in the baselines and the number of foci k is set as 50. For initial evaluation of the framework, we choose α=1 and β=1. The performance for different values of α and β will be presented in future subsections.

Performance Evaluation

The results of the evaluations are presented in Table 2. From Table 2, we can see that the proposed framework has outperformed the baselines by a considerable margin. We conducted a paired t-test to compare the performance of our framework with that of the baselines, and the results indicated the difference between them is significant. We make the following observations from the table.

The proposed framework gives more than 300% improvement over random selection. We can see that simple formulations like the one in Aardvark that considers social network information performs on par with complex topical models using only content such as STM. The proposed framework also performs significantly better than methods identifying subject matter experts as answerers such as TSPR. This emphasizes the importance of social context to identify answerers to social media questions.

Considering shared foci between the asker and the answerer improves the performance over methods like Aardvark not utilizing community memberships. This shows the effectiveness of using social foci to exploit social context. Incorporating question information to consider the overlap only in the foci related to the question gives further improvement in the performance.

In summary, by designing approaches based on shared social context and exploiting the structure of social ties, the proposed framework can effectively identify answerers for social media questions in the dataset. Next, we wish to understand the effect of content and network information on the performance of our framework

Effect of Content and Network Information

In the model presented in Eq. (3), α and β control the proportion of the network and content information respectively.

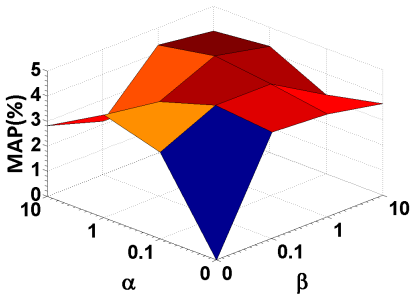


Figure 2: Effect of variation of content and network proportions on the framework performance for MAP.

In order to evaluate the framework for different proportions of content and network, we set $\alpha = [0.1, 1, 10]$ and $\beta = [0, 0.1, 1, 10]$ and plot the values for MAP in Fig. 2 arbitrarily using cosine similarity as the similarity metric. We make the following observations from the figure.

A general trend in Fig. 2 is a peak at the main diagonal of the α and β axes and an off-diagonal dip. This shows that the framework works best for nearly equal proportions of network and content information. The MAP value is greater than 3% for all α and β except for low proportions of network information ($\alpha = 10$, $\beta = [0, 0.1]$). This emphasizes the importance of social connections of the asker for identifying answerers to social media questions. The lowest performance across all parameter values is more than twice than random ordering indicating the effectiveness of the framework for low relative proportions of content or network information. Overall, the MAP value is above 3% for different combinations of α and β indicating the effectiveness of the framework for a wide range of parameter values.

In summary, the framework performs well over different proportions of network and content and is robust to their variation. An appropriate combination of network and content information can optimize the effectiveness of the framework for identifying answerers to social media questions.

Performance across Question Categories

Literature on social media questions has identified kinds of questions people ask on Twitter. Recommendation, opinions, factual and rhetorical questions are popular questions asked on Twitter (Morris, Teevan, and Panovich 2010; Paul, Hong, and Chi 2011). We select four categories related to subjective questions, "Suggestions", "Opinion", "Favor", and "Rhetorical", and evaluate our framework in identifying answerers for different question categories.

We employed human labelling to assign category labels to questions. Three people independently labeled the questions, and the labels were assigned using majority selection. Employing this procedure, 93.5% of the questions were assigned to either of the four categories and the framework was evaluated on them. The results of the evaluations are presented in Table 3. The distribution of different question categories is given in the first column. The performance for different categories is listed in the other columns. The improvement over (Horowitz and Kamvar 2010), the nearest baseline not a part of our method, for different question categories is shown in the brackets.

Categories	Parts	MRR	MAP@5
Suggestions	39.83%	4.27%(+2.23%)	4.68%(+1.78%)
Opinion	16.42%	2.67%(+1.43%)	2.38%(+1.61%)
Favor	30.51%	3.65%(+1.55%)	4.39%(+1.01%)
Rhetorical	6.74%	1.75%(+1.17%)	0%(+0%)

Table 3: Performance for different question categories.

From the table, we see that the framework gives considerable improvements over all the selected question categories. A paired t-test suggested that the improvements are significant indicating that the framework is effective in finding answerers to a wide range of question categories in Twitter. The best performance can be seen in "Suggestions" and "Favor" categories and the performance in "Opinions" is relatively lower. These results suggest that identifying answerers for the "Opinion" category might depend on additional factors such as similarity of views in a given topic. The framework gives the lowest performance for questions in the "Rhetorical" category. Rhetorical questions are classified as conversational questions in the literature (Harper, Moy, and Konstan 2009). They might be used as an expression of opinion or to initiate a conversation and not to express an information need.

Conclusion and Future Work

Online social media provides a new platform for people seeking information from their social connections. Social media questions represent a form of information seeking behavior of users. Questions are subjective and personal to the asker, and his social context is useful to identify answerers. We draw from sociological theories to present a novel framework to infer the shared context between the asker and the answerers in the question domain. We evaluate the framework on questions on Twitter and demonstrate its effectiveness in identifying answerers. The framework is robust to a wide range of proportions of network and content information and categories of social media questions. The paper provides the first framework with experimental evaluations to identify answerers to questions in social media.

Frameworks exist to identify answerers to factual questions prevalent in community Q&A platforms like Yahoo Answers and StackOverflow. Incorporating concepts from them will enable us to tackle more diverse questions. During situations like natural disasters, social media users propagate requests for help throughout the network. Identifying answerers in these situations will require an understanding of information propagation and information seeking behavior. Identifying users providing misinformation to questions in social media will help to increase the effectiveness of social media as a quality information source.

Acknowledgements

This material is based upon work supported by, or in part by, Office of Naval Research (ONR) under grant number N000141010091.

References

Adamic, L. A.; Zhang, J.; Bakshy, E.; and Ackerman, M. S. 2008. Knowledge sharing and yahoo answers: everyone knows something. In *WWW*.

Bian, J.; Liu, Y.; Agichtein, E.; and Zha, H. 2008. Finding the right facts in the crowd: factoid question answering over social media. In *WWW*.

Bozzon, A.; Brambilla, M.; Ceri, S.; Silvestri, M.; and Vesci, G. 2013. Choosing the right crowd: expert finding in social networks. In *EDBT*. ACM.

Burt, R. S. 2009. *Structural holes: The social structure of competition*. Harvard university press.

Case, D. O. 2012. *Looking for information: A survey of research on information seeking, needs and behavior*. Emerald Group Publishing.

Cong, G.; Wang, L.; Lin, C.-Y.; Song, Y.-I.; and Sun, Y. 2008. Finding question-answer pairs from online forums. In *SIGIR*.

Ding, C.; Li, T.; Peng, W.; and Park, H. 2006. Orthogonal nonnegative matrix t-factorizations for clustering. In *KDD*.

Ding, C.; Li, T.; and Jordan, M. I. 2008. Nonnegative matrix factorization for combinatorial optimization: Spectral clustering, graph matching, and clique finding. In *ICDM*.

Du, L.; Buntine, W.; and Jin, H. 2010. A segmented topic model based on the two-parameter poisson-dirichlet process. *JMLR*.

Efron, M., and Winget, M. 2010a. Questions are content: a taxonomy of questions in a microblogging environment. *ASIST*.

Efron, M., and Winget, M. 2010b. Questions are content: A taxonomy of questions in a microblogging environment. *ASIST*.

Ellison, N. B.; Gray, R.; Vitak, J.; Lampe, C.; and Fiore, A. T. 2013. Calling all facebook friends: Exploring requests for help on facebook. In *ICWSM*.

Feld, S. L. 1981. The focused organization of social ties. *AJS*.

Gray, R.; Ellison, N. B.; Vitak, J.; and Lampe, C. 2013. Who wants to know?: question-asking and answering practices among facebook users. In *CSCW*.

Harper, F. M.; Moy, D.; and Konstan, J. A. 2009. Facts or friends?: distinguishing informational and conversational questions in social q&a sites. In *CHI*.

Hecht, B.; Teevan, J.; Morris, M. R.; and Liebling, D. J. 2012. Searchbuddies: Bringing search engines into the conversation. *ICWSM*.

Horowitz, D., and Kamvar, S. D. 2010. The anatomy of a large-scale social search engine. In *WWW*.

Jeong, J.-W.; Morris, M. R.; Teevan, J.; and Liebling, D. 2013. A crowd-powered socially embedded search engine. *ICWSM*.

Jurczyk, P., and Agichtein, E. 2007. Discovering authorities in question answer communities by using link analysis. In *CIKM*.

Kumar, S.; Morstatter, F.; and Liu, H. 2013. *Twitter Data Analytics*. Springer.

Lampe, C.; Gray, R.; Fiore, A. T.; and Ellison, N. 2014. Help is on the way: Patterns of responses to resource requests on facebook. In *CSCW*. ACM.

Lee, D. D., and Seung, H. S. 2000. Algorithms for non-negative matrix factorization. In *NIPS*.

Lee, U.; Kang, H.; Yi, E.; Yi, M.; and Kantola, J. 2012. Understanding mobile q&a usage: An exploratory study. In *CHI*.

Marsden, P. V. 1988. Homogeneity in confiding relations. *Social networks*.

Morris, M. R.; Teevan, J.; and Panovich, K. 2010. What do people ask their social networks, and why?: a survey study of status message q&a behavior. In *CHI*.

Nandi, A.; Paparizos, S.; Shafer, J. C.; and Agrawal, R. 2013. With a little help from my friends. In *ICDE, 2013*.

Pal, A., and Counts, S. 2011. Identifying topical authorities in microblogs. In *WSDM*. ACM.

Panovich, K.; Miller, R.; and Karger, D. 2012. Tie strength in question & answer on social network sites. In *CSCW*.

Paul, S. A.; Hong, L.; and Chi, E. H. 2011. Is twitter a good place for asking questions? a characterization study. In *ICWSM*.

Radev, D. R.; Qi, H.; Wu, H.; and Fan, W. 2002. Evaluating web-based question answering systems. *Ann Arbor*.

Riahi, F.; Zolaktaf, Z.; Shafiei, M.; and Milios, E. 2012. Finding expert users in community question answering. In *WWW companion*.

Seung, H., and Lee, D. 2001. Algorithms for non-negative matrix factorization. *NIPS*.

Tang, L., and Liu, H. 2009. Relational learning via latent social dimensions. In *KDD*.

Wang, G.; Gill, K.; Mohanlal, M.; Zheng, H.; and Zhao, B. Y. 2013a. Wisdom in the social crowd: an analysis of quora. In *WWW*.

Wang, Y.; Wang, L.; Li, Y.; He, D.; Liu, T.-Y.; and Chen, W. 2013b. A theoretical analysis of ndcg type ranking measures. *arXiv preprint arXiv:1304.6480*.

Yang, J.; Morris, M. R.; Teevan, J.; Adamic, L. A.; and Ackerman, M. S. 2011. Culture matters: A survey study of social q&a behavior. *ICWSM*.

Yang, L.; Qiu, M.; Gottipati, S.; Zhu, F.; Jiang, J.; Sun, H.; and Chen, Z. 2013. Cqarank: jointly model topics and expertise in community question answering. In *CIKM*, 99–108. ACM.

Yang, J.; McAuley, J.; and Leskovec, J. 2013. Community detection in networks with node attributes. *ICDM*.

Zhou, G.; Lai, S.; Liu, K.; and Zhao, J. 2012. Topic-sensitive probabilistic model for expert finding in question answer communities. In *CIKM*.

Zhu, H.; Chen, E.; Xiong, H.; Cao, H.; and Tian, J. 2013. Ranking user authority with relevant knowledge categories for expert finding. *WWW*.