SWAT: Social Web Application for Team Recommendation

Stefano Braghin, Jackson Tan Teck Yong
Nanyang Technological University, Singapore
{s.braghin,jacktty}@ntu.edu.sg

Anthony Ventresque
University College Dublin, Ireland
anthony.ventresque@ucd.ie

Anwitaman Datta
NTU Singapore
anwitaman@ntu.edu.sg

***Abstract*—Team recommendation aids *decision support*, by not only identifying individuals who are experts for various aspects of a complex task, but also determining various properties of the team as a group. Several aspects such as cohesion and repetition of teams have been identified as important indicators, besides individuals' expertise, in determining how well a team performs. While such information often do not exist explicitly, *digital footprint* of users' activities can be harnessed to retrieve the same from diverse sources. In this work, we lay out a proof-of-concept on how to do so in the case of scientific knowledge workers, as well as demonstrate some necessary visualization, manipulation and communication tools to determine and manage multi-disciplinary teams. While the focus of our presentation is the specific application 'SWAT' for team recommendation, it also serves as a vehicle demonstrating how, in general, apparently disparate data sources can be harnessed to provide decision support guided by suitable analytics.**

***Keywords* —** recommendation system, decision support, visualization, team, social networks, analytics

I. Introduction

Note: SWAT is an actually prototyped system, which will be demonstrated. More information can also be found at http://sands.sce.ntu.edu.sg/SWAT

The old adage goes *united we stand, divide we fall*. 'Team work' has become increasingly important for success in many walks of life, particularly so for knowledge workers. Many outstanding and emerging scientific and technical challenges are often complex, at the interface of multiple disciplines. Furthermore, frequently they become 'hot topics' pursued by many competitors. To be able to solve the problem at all, and do so within a competitive time-window, forming a 'good team' with expertise in the various facets of the problem is crucial.

In fact, the problem of *multidisciplinary team recommendation* itself is a ready example of such a multifaceted problem - needing expertise retrieval [6], [8] (which in turn rely on *topic extraction*, *influence network analysis*, etc.), analysis of *team cohesion* and dynamics to determine important parameters of a team's formation and success [3], [7], [11], [12], [13] and *information extraction* from online sources [10], to name a few.

This work was funded by A*STAR SERC TSRP Grant 1021580038. Anthony Ventresque contributed in the development of SWAT when he was working at NTU Singapore as a research fellow. Contact author: Anwitaman Datta.

In the past, the problem was arguably intractable due to lack of data, and researchers mostly relied on costly and time-consuming individual surveys [1]. Additionally, there were no suitable platform for large-scale deployment of a team exploration and recommendation system. In contrast, at present, data is often readily available on the social web, and the challenge is rather to discern the useful 'information' from the 'noise', and cumulate apparently disparate data to make sense of the big picture. In that sense, while the scope of the presented 'SWAT' system is limited, and we accordingly restrict rest of our discussions as well, one can also view SWAT as one out of many possible examples, where abundance of digital footprints are harnessed with suitable analytics to aid decision support.

The rest of our discussions will be specific to the application of team recommendation, and we will not delve on the issue of big-data analytics directly. However, without crucial enablers such as Web 2.0 online data repositories & social-networks and the analytics of big data generated as digital footprint of individuals' activities online, a SWAT like system could not possibly be conceived, let alone realized.

Reverting back to the specific problem under discussion, an obvious challenge could be to 'find the best team'. We argue that such a quest is however neither tractable, nor relevant. One might for instance consider a retro-prediction testbed (eg. published papers) and check our system against it. But real teams are somewhat unpredictable, or driven by so many parameters we do not have our hands on (eg. occasional informal discussions in conference, or private email exchanges not captured in the readily available digital foot prints) that it sounds intractable.

More crucially, we will like to emphasize that a 'recommendation' based decision support system is not about prediction: instead it means making propositions that eases the end user's decision process, possibly by readily providing 'partial solutions' or making it easy to explore the solution space. Our heuristics however build upon a lot of rigorous works solving individual sub-problems such as *topic extraction* - both in general, as well as for the specific purpose of identifying experts, and on understanding *team dynamics*, making SWAT a robust team recommendation system.

Motivation: While the degree of promiscuity is naturally increasing in scientific work, for instance evidenced by studies observing that in the academic world, the number of authors

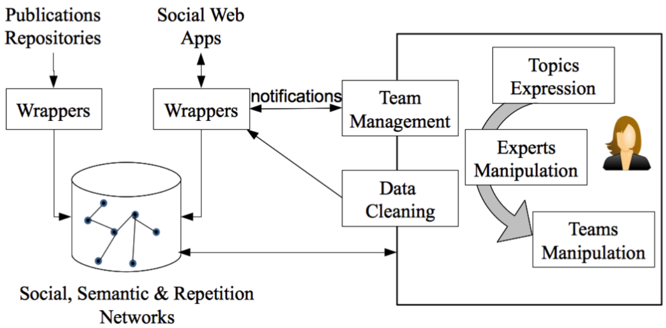

Fig. 1. SWAT system overview

per paper and the number of coauthors per author are increasing [4], [9], people primarily tend to work with the same set of personal acquaintances but miss new potential collaborators (and as a result, opportunities), since people are generally barely aware of experts from various other topics involved to carry out a complex multidisciplinary work. Thus, at present, often new collaborations mainly happen out of chance meetings, or based on chance personal recommendations from common acquaintances. Yet, there is also a growing realization that working at the boundary of distinct disciplines have the potential to solve problems that may be more difficult, if not impossible - to deal with without tools and knowledge from the multiple domains. Thus, a team recommendation system bridges the gap between the typical modus operandi, where collaborations are ad-hoc, and the need to find experts from diverse domains, while also trying to ensure that the resulting team members would share a good rapport to work together. Given the subjectivity of the nature of the problem, as well as the impossibility to acquire and analyze all possible data - trying to find an absolute and objective answer is not meaningful. However, providing reasonable recommendations has become particularly feasible given the advent of online data repositories and social networks.

The social web has emerged as a disruptive technology, changing the way many things are done. This holds true also for finding new jobs, head hunting, etc., LinkedIn being the most prominent social networking site enabling such activities. An expert and team recommendation system pushes the envelope further by 'automating' the social and expertise network exploration process, and thus providing a means to zero-in on the pertinent 'needle' out of the 'haystack' (information overload/noise). While finding experts for one specific domain is in itself non-trivial, finding a group of experts, possibly from different domains, and analyzing which combinations have what social relations is even more so.

To that end, in this paper, we go a step beyond existing efforts [2], and address several problems of team recommendation, leveraging on which, we have built a proof of concept web application called SWAT (Social Web Application for Team Recommendation).[1] The current proof of concept of SWAT is based on scientific knowledge workers, and primarily leverages on openly available information. However, the implementation is modular (see Figure 1), and the basic framework can also be readily deployed in diverse other scenarios, such as, based on open source software history, user declared information (e.g., what LinkedIn users often provide) or in enterprise intranet, while leveraging on other necessary/relevant modules that we present.

Specifically, in the current SWAT instantiation we (i) perform data extraction from online repositories (such as DBLP, Elsevier, ...), (ii) leverage on academics' technophilia to collect reliable friendship connections, while (iii) cleaning the data (for instance, without the burden of filling multiple forms on the user's part, to); (iv) extract topics and disambiguate names, which all lead to the (v) construction of data suitable for team recommendation algorithms.

SWAT builds on, but significantly extends our previous work T-RecS [2] in several ways. Firstly, T-RecS was based on a well defined and static dataset (NTU's internal publication records), while SWAT has more generic inputs, which in turn required new modules for data cleaning and disambiguation. SWAT can thus also readily incorporate new data sets, which however also means that the data to be analyzed may be *dynamic and continuously growing*. SWAT uses improved team recommendation algorithms (based on the new *multi-dimensional social network* information representation used in the backend), and also provides *enhanced functionalities* in terms of team recommendation and exploration, providing the user with several *visualization*s of recommended teams, as well as a *notification mechanism* with a semi-automatic update of teams to deal with individuals' availability and willingness to participate into teams selected by managers. While SWAT is an independent web application, it is also *integrated with Facebook*, which allows some additional functionalities: ease of finding other users and additional dimension of social network (which may not be captured in other datasets used: e.g., a colleague or friend who has never been a coauthor) and facilitate active interactions with such users using Facebook's existing communication tools. Such features make SWAT an useful decision support system and collaboration enabler. For conciseness and specificity, the rest of the paper describes the details and datasets used in the current SWAT implementation - without necessarily outlining how the ideas can be generalized for other settings. Acquiring, storing, matching and analyzing the huge corpus of dynamic social network data from disparate sources provide a lot of data management and computing challenges as well, however, the scope of this presentation will exclude the general nuances of managing such big data, and instead we will focus primarily on the demonstrating SWAT application.

[1]Some more information on the application can be found at http://sands.sce.ntu.edu.sg/SWAT

II. HARVESTING & CLEANING DATA

An important information needed for team recommendation consists of interactions between people and concepts. For scientific knowledge workers, there are numerous online repositories (both open as well as controlled) which can be harnessed. In a previous work [13] we considered three projections of the digital information: semantic network, social network and socio-semantic co-occurrences; and showed that all of them are of importance for team formation, since they give an accurate set of information for representing expertise, cohesiveness and team and concept repetitions, which are in turn four important features of team quality and dynamics.

Raw data in our prototype comes from online repositories. Specifically, data from DBLP[2] was used to fetch information about author names, publications titles, venues and links to electronic editions. We accordingly crawled abstract of the publications hosted by other, often non-open, repositories. We capture data regarding who works on what topics, with who all and how often. Thus other information repository, including from other walks of life (say, software development teams) - can be plugged into SWAT.

Two principal problems when integrating data from the 'wild' are *identification of concepts* of the collaboration and *people disambiguation*. Concept extraction consists in automatically finding in a document a set of concepts that best describes it. There are various techniques to do so, for instance by extracting concepts from the document (sometimes with the help of external resources to select the most representative keywords) or assigning concepts from external vocabularies (controlled or not). In SWAT we employ a mixed approach: assignment of concepts from Wikipedia according to what we find in the document, and extraction of the best ones according to relations in the Wikipedia hierarchy. This technique uses keyword disambiguation [5] first to map words in the document to Wikipedia pages (either articles or categories); these words are then clustered according to their similarity value in Wikipedia hierarchy and the system extracts a candidate from each cluster; top candidates are eventually selected to represent the document. Person name disambiguation is a very hard and open problem in its own right, so we use some very preliminary heuristics at the moment (which can be replaced by any state-of-the-art solutions in future). There are dual aspects to the name disambiguation problems, namely mapping different named entities to the same person and split same name to different people, SWAT relies on names of people together with topics and neighbors (friends/colleagues) to compute similarity value between names, and merge them if the value reaches a certain threshold. The odds are good that two people with very similar names, same research topics and friends are the same. On the other hand, we compute a person cohesion value which reflect the fact that various interactions of this person are more likely than random. If it appears that a person work with two different group of people and on two different set of topics throughout time, that could mean that there are in fact two distinct persons.

Given the prematurity of techniques solving the disambiguation problems, and the simplicity of the heuristics currently applied in SWAT, our data is expected to still be dirtier than what can be wished for, i.e. some topic or name assignments to people are incorrect. Manual check for millions of academics and tens of million of publications is infeasible. However, one can rely on crowdsourcing the task. SWAT allows its users to correct mistakes and add missing information, such as fixing misspelled paper titles, adding missing authors or the merging of academic profiles divided because of the use of aliases (see Section IV for details).

Data integrity issues arise from granting such power to the users, but security issues are out of the scope of the current implementation, where we assume that users provide correct data and do not behave maliciously to pollute the data. We plan to enhance SWAT in future by introducing security checks to mitigate vandalism.

III. TEAM FORMATION & MANAGEMENT

SWAT extends the ideas from T-RecS [2] to recommend teams, and gives a list of possible teams according to a trade-off between four important features: expertise, social cohesiveness, team repetition and mutual interest. These features were proposed and explored in [3], [4], [13].

In T-RecS, we presented an algorithm based on expertise and cohesiveness (which continue to be contributing factors in SWAT). Expertise value of a team is a combination of expertise values of team members; maximum, minimum or average, to name few of the possible combinations. Each one has different impact on team formation, e.g., to have a good expert for each skill, to avoid weak links or to balance strong and not so strong members in the team. On the other hand, cohesiveness reflects how close team members are to each other. A proximity value between any two people is computed on the social graph, determining in turn a clustering coefficient of the team to quantify cohesiveness.

SWAT's recommendation algorithm is based on our previous empirical study [13] showing that, besides expertise and cohesiveness, concept and people repetition are important factors. Team repetition refers to recurrence of a subset of team members in previous teams. It is significantly different from cohesiveness as it does not consider one-to-one relationships but group combination. Namely, we define two new metrics to evaluate potential teams: *user team repetition* and *concept team repetition*.

Let $\mathcal{T} = \{t_1, \ldots, t_n\}$ be the set of past teams, with $t_i = (U, C)$ where U is the set of individual users and C is the set of concepts. Moreover, given a team $t_i \in \mathcal{T}$, U^{t_i} denotes the set of users and C^{t_i} the set of concepts associated with t_i.

The *user team repetition* of a given team t counts the occurrences of subsets of its members as members of past teams. We denote such metric with $\mathcal{U}(t)$. Note that, we restrict the cardinality of such subsets to be greater than two (since we do not consider a single user as a team). See Eq. 1, where

[2]http://www.informatik.uni-trier.de/~ley/db/

the function $\phi(\hat{U},\mathcal{T})$ counts the number of teams $t' \in \mathcal{T}$ with $\hat{U} = U^{t'}$.

$$\mathcal{U}(t) = \sum_{\hat{U}\in\mathcal{P}(U^t) \text{ with } |\hat{U}|\geq 2} \phi(\hat{U},\mathcal{T}) \qquad (1)$$

On the other hand, mutual interest – or conceptual repetition – considers reiteration of the a significant subset of concepts and of academics from potential team in previous collaborations. Following the previous ideas, Eq 2 checks whether subsets of academics worked together on a significant subset of concepts/skills required in the team. Hence, *concept team repetition*, denoted by $\mathcal{C}(t)$, considers the teams composed of members of t and their repetition of concepts of t. Such metric measures the similarity of the 'scope' – defined as required concepts – of the evaluated team t with the past teams. This is achieved by finding, for each subset of the teams in t, the maximum number of similar concepts in past teams.

$$\mathcal{C}(t) = \sum_{\hat{U}\in\mathcal{P}(U^t)\setminus\varnothing} \max_{\hat{t}\in\mathcal{T} \text{ s.t. } U^{\hat{t}}=\hat{U}} \left|C^t \cap C^{\hat{t}}\right| \qquad (2)$$

SWAT allows users to customize the importance of these four factors specifying, for each of them, a weight which is used during the ranking of the teams. SWAT provides a set of default weights that users may easily modify in order to evaluate the outcome of different configurations of the ranking mechanism. Beside that, SWAT allows to specify further constraints to both the selection of the experts and the classification of the teams such as geographic/affiliation diversity/affinity and conflict of interests.

IV. User Experience

User interfaces are key elements of SWAT. They need to be simple, not only the steps of team features selection, navigation, etc. to simplify the learning curve, but also to leverage on users' knowledge and motivation to clean and improve the data.

SWAT can be used either through a social web application (current implementation is integrated with Facebook only (see Figure 2), but it can also be academia, Scispace, LinkedIn, EDAS, etc.) or without it as a stand alone service. In the later case SWAT still provides all the functionalities for team recommendation, but does not support team management, such as notification to team members, or harnessing extra social network information. In contrast, if a user registers with SWAT in a social web application then she is required to grant some permissions to the application, for the application to be able to access the user's data stored in the social network. SWAT then asks her to map the social network's account with academic(s) from our database. There is no disclosure of identity at this point as mapping stays private. But the system uses this first step to correct some elements of our database: SWAT presents the information we have ('friends' that can be potential collaborators, publications, etc.) and the user can add/remove some information and fix the remaining disambiguation false positives/negatives.

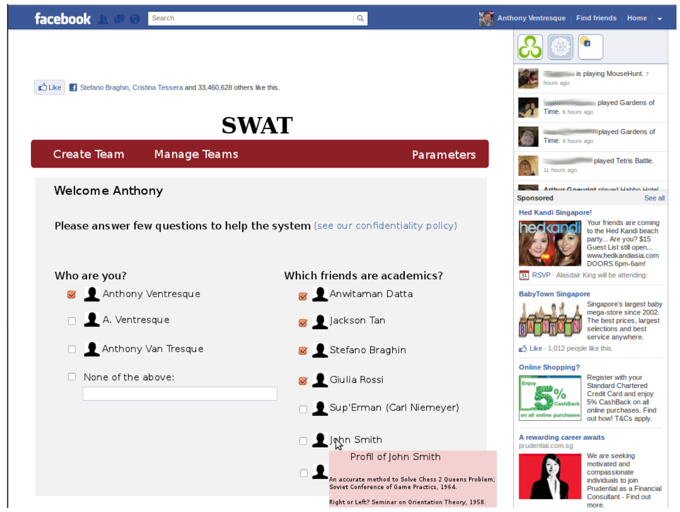

Fig. 2. SWAT's registration page within Facebook

The creation of a new team in SWAT is a three step process beginning when the user clicks on 'Create team': (i) user selects the skills required by the team, (ii) visualizes and modifies lists of experts given by SWAT and (iii) navigates the generated teams. Navigation and visualization of the multi-dimensional social network and potential teams are important for usability. Therefore, teams are presented according to a ranking, but the user can customize the metrics to re-rank, and also select other visualizations (see Figure 3). The current SWAT implementation supports three kinds of visualizations, and allows modular integration of others. Besides the traditional list of top teams implemented in T-RecS, SWAT's users can display any team either as a radar chart (see Figure 4), as a graph of academics and concepts (see Figure 5), or automatically generate a short paragraph summarizing the team's vitae by clicking on the corresponding icon (A in Figure 3). Radar chart dimensions include cohesion, expertise, sociability (global number of collaborators of team members), activity (global number of publications of team members), productivity (fractional count of publications per team member), impact of publications (ratio of activity and importance of publications). Graph displays academics, their social links, their competencies regarding team's needs, and some other useful information. Textual summary is a short paragraph with main characteristics of team members: their common history, the members profile and achievements (in terms of number or quality of publications, topics), etc.

Users can also edit some parameters (e.g., add/remove members manually) and see the list being modified consequently (B in Figure 3).

When an user selects a team – clicking on the mark icon (C in Figure 3) – she may wish to contact the potential members to ask for their availability for, and coordinate the activities of a project. SWAT utilizes notification service of social web application(s) – or any other available communication mechanism such as e-mail – to automatically contact those among the team members registered with it/them. Once any social web application user receives a notification she can either fully or conditionally accept, or reject. The original user receives these responses and can modify the team, aided again by SWAT to find new members if needed.

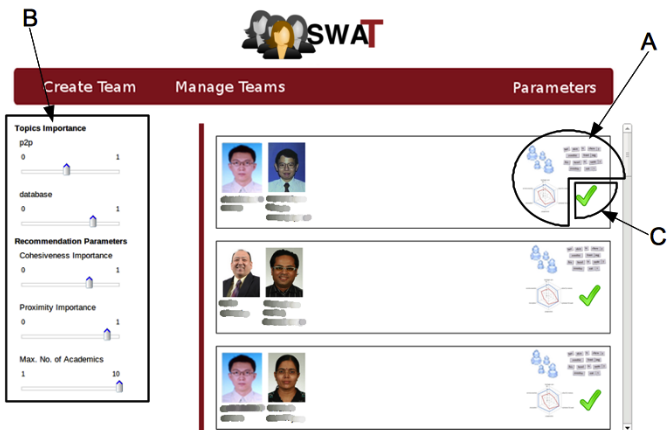

Fig. 3. Navigation and manipulation of recommended teams.

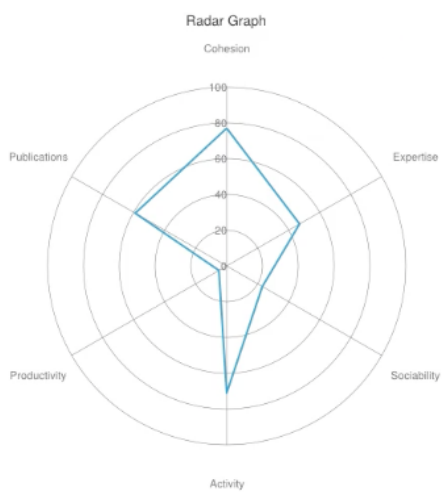

Fig. 4. An example of radar chart

Once a team is selected, and depending on the user's registration with SWAT social web application, collaborators who are also registered will automatically receive a notification. They can accept, refuse, conditionally accept and give some feedback to the manager (A in Figure 6). SWAT automatically notifies the manager and recommends new collaborators if needed. One section of the management board is accessible to all team members (B in Figure 6): they can message other members and synchronize jobs/meetings.

V. Demonstration

In this demo we will show: (i) the power of data cleaning (ii) a real application of a team recommendation (iii) the flexibility of our system according to user's needs (iv) subsequent management of selected teams.

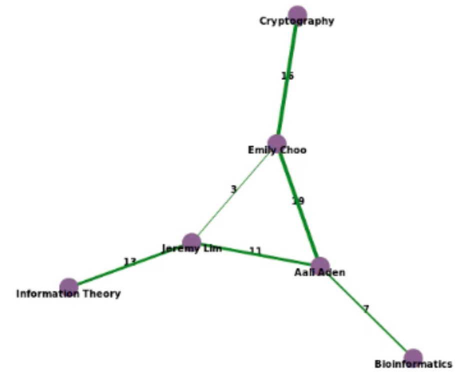

Fig. 5. An example of team graph

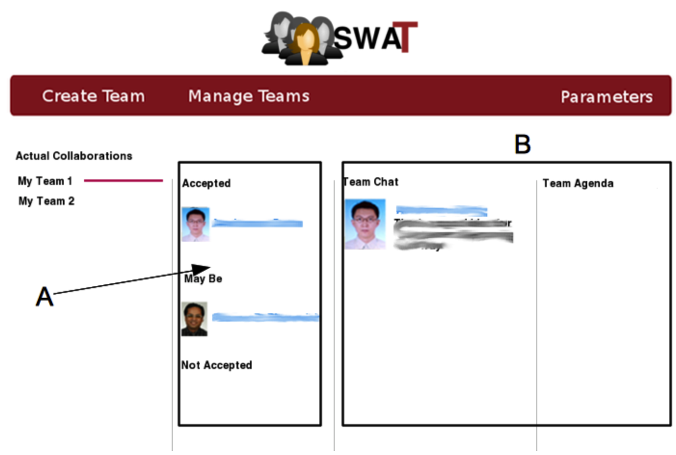

Fig. 6. Management of collaborators availability and interactions between team members.

We will play out two scenarios: an ego-centric search personalized to user's social network and preferences and a manager looking for a scientific team able to carry out a project.

The first scenario considers a user registering into our social web application, handily matching her profile to what we already get in our database and cleaning it easily: checking if the profile is correct, and giving more information about her friends. She will then create a team to fulfill some specific requirements and the system will react according to her preferences, social network, etc. This team can then be managed through the social web application.

We will then play out a manager who wants to get advised from our web application, without registration onto the social application. She will describe the skills required in a team, and the system will propose candidate teams that she can navigate and manipulate.

Anchored around this demonstration, we will also enumerate some of the encountered as well as open challenges in managing the huge amount of dynamic data from disparate sources.

DISCLAIMER

These algorithms, some of them obtained from the literature and some others developed by ourselves, are possibly far from perfect, and the research community as a whole continuously continues to refine them. Likewise, the data-sets we use are not complete. Consequently, the scores - both on individuals' competence on a topic as well as social cohesiveness should not be considered as a reflection of the ground truth, and are instead used just for the purpose of illustration.

REFERENCES

[1] M. Case, Contractor, J. N., P. M., S. C.-Y. Lu, and B. J. O'Keefe. Team engineering analysis and modeling: Toward a normative model of team interaction. In *SMC*, 1992.

[2] A. Datta, J. Tan Teck Yong, and A. Ventresque. T-RecS: Team recommendation system through expertise and cohesiveness. In *WWW*, 2011.

[3] M. Geipel. Dynamics of communities and code in open source software. *ETH Doctoral Thesis*, (18480), 2009.

[4] J. W. Grossman. The evolution of the mathematical research collaboration graph. In *Congressus Numerantium*, 2002.

[5] C. Li, A. Sun, and A. Datta. A generalized method for word sense disambiguation based on wikipedia. In *ECIR*, 2011.

[6] C.-Y. Lin, N. Cao, S. X. Liu, S. Papadimitriou, J. Sun, and X. Yan. Smallblue: Social network analysis for expertise search and collective intelligence. *ICDE*, 2009.

[7] K. Luther and A. Bruckman. Leadership and success factors in online creative collaboration. *Potentials, IEEE*, 30(5), 2011.

[8] D. W. McDonald and M. S. Ackerman. Expertise recommender: a flexible recommendation system and architecture. In *CSCW*, 2000.

[9] M. Newman. Who Is the Best Connected Scientist? A Study of Scientific Coauthorship Networks. *Complex Networks*, 2004.

[10] S. Sarawagi. Information extraction. *Found. Trends databases*, 1:261–377, March 2008.

[11] K. J. Shim, S. Damania, C. DeLong, and J. Srivastava. Player and team performance in everquest ii and halo 3. *Potentials, IEEE*, 30(5), 2011.

[12] P. Turek, A. Wierzbicki, R. Nielek, and A. Datta. Wikiteams: How do they achieve success? *Potentials, IEEE*, 30(5), 2011.

[13] A. Ventresque, J. Tan Teck Yong, and A. Datta. Impact of Expertise, Social Cohesiveness and Team Repetition for Academic Team Recommendation. In *Socinfo*, 2011.