Composing near-optimal expert teams: a trade-off between skills and connectivity

Christoph Dorn and Schahram Dustdar

Distributed Systems Group, Vienna University of Technology, 1040 Vienna, Austria,
lastname@infosys.tuwien.ac.at,
WWW home page: http://www.infosys.tuwien.ac.at/staff

Abstract. Rapidly changing business requirements necessitate the ad-hoc composition of expert teams to handle complex business cases. Expert-centric properties such as skills, however, are insufficient to assemble an effective team. The given interaction structure determines to a large degree how well the experts can be expected to collaborate. This paper addresses the team composition problem which consists of expert interaction network extraction, skill profile creation, and ultimately team formation. We provide a heuristic for finding near-optimal teams that yield the best trade-off between skill coverage and team connectivity. Finally, we apply a real-world data set to demonstrate the applicability and benefits of our approach.

Keywords: social network, team formation, simulated annealing, skill connectivity tradeoff

1 Introduction

Over the past years we have observed a trend towards online knowledge creation and sharing (e.g., Slashdot[1], Yahoo! Answers[2]). People increasingly apply their expertise online to answer other users' questions or provide additional information on topics under discussion. Rapidly changing business requirements keep individual companies from employing a large set of experts that continuously cover the required skill set. Exploration of online communities allows dynamic access to the top experts of the desired expertise.

Previous work focused on identifying the most important experts in an online community (e.g., [**?**]). Complex business cases, however, require the complementary expertise of multiple experts that need to collaborate closely. A team of top experts will be most effective if they have interacted before and thus exhibit confidence in each other's expertise. The problem is finding the best trade-off between maximum skill coverage and maximum interaction connectivity. Finding an optimal team configuration is non-trivial as the search space grows exponentially with the number of required skills and available experts.

[1] http://slashdot.org/
[2] http://answers.yahoo.com/

In this paper we present a mechanism for extracting an expert network and corresponding expert skill profiles from online discussion threads. Our novel trade-off model allows fine-grained preference configuration of team connectivity over maximum skill coverage. We provide a heuristic to extract the optimum team composition from an expert network for a given trade-off configuration. As the skill data originates from discussion sites, we envision the resulting team compositions to collaborate well over the internet rather than work face to face.

Section **??** compares our work to previous research efforts. Section **??** outlines our approach in more detail based on a motivating example. Subsequently, Section **??** describes the mechanism for extracting an expert network and corresponding skill profiles from discussion threads. Section **??** provides the formal definition of the team formation problem. Section **??** demonstrates the adaptation of Simulated Annealing to our problem. The evaluation in Section **??** applies a real-world data set to demonstrate the effectiveness and efficiency of our approach. The paper concludes with an outlook on future work.

2 Related Work

Team formation is an intensely studied problem in the operation research domain. Most approaches model the problem as finding the best match of experts to required skills taking into account multiple dimensions from technical skills, cognitive properties, and personal motivation [**?**,**?**,**?**]. Such research focuses only on properties of individual experts that are independent of the resulting team configuration.

Recent efforts introduce social network information to enhance the skill profile of individual members. Hyeongon et al. [**?**] measure the *familiarity* between experts to derive a person's *know-who*. Cheatham and Cleereman [**?**] apply social network analysis to detect common interests and collaborations. The extracted information, however, is applied independently from the overall team structure. These mechanisms present opportunities for refinement of the skill modeling and configuration aspects of our approach but remain otherwise complementary.

To the best of our knowledge, Theodoros et al. [**?**] discuss the only team formation approach that specifically focuses on the expert network for determining the most suitable team. Our approach differs in two significant aspects. First, we model a trade-off between skill coverage and team connectivity whereas [**?**] treats every expert above a certain skill threshold as equally suitable and ignores every expert below that threshold. Second, our algorithm aims for a fully connected team graph (i.e., relations between every pair of experts). Theodoros et al. optimize the team connectivity based on a minimum spanning tree (MST). We argue that it is more important to focus on having most members well connected (i.e. everybody trusts (almost) everybody else) within the team than focusing only on the strongest ties within the team.

Analysis of various network topologies [**?**] has demonstrated the impact of the network structure on efficient team formation. General research on the formation of groups in large scale social networks [**?**] helps to understand the involved dy-

namic aspects but does not provide the algorithms for identifying optimal team configurations. Investigations into the structure of various real-world networks provides vital understanding of the underlying network characteristics relevant to the team formation problem [**?**,**?**]. Papers on existing online expert communities such as Slashdot [**?**] and Yahoo! answers [**?**] yield specific knowledge about the social network structure and expertise distribution that need to be supported by a team formation mechanism. Complementary approaches regarding extraction of expert networks and their skill profile include mining of email data sets [**?**] or open source software repositories [**?**].

Related research efforts based on non-functional aspects (i.e., non-skill related aspects) can also be found in the domain of service composition [**?**]. Here, services with the required capabilities need to be combined to provide a desirable, overall functionality. Composition (i.e., formation) is driven by the client's preferences [**?**], environment context [**?**,**?**], or service context (i.e., current expert context) [**?**]. We can take inspiration from such research to refine the properties and requirements of teams to include context such as expert availability or location. Nonetheless, the network structure remains equally unexplored in service composition.

In contrast, the network structure has gained significant impact for determining the most important network element. A prominent example of a graph-based global importance metric is Google's page rank [**?**]. An extended version [**?**] yields total ranks by aggregating search-topic-specific ranks. Inspired by the page rank algorithm, Schall [**?**] applies interaction intensities and skills to rank humans in mixed service-oriented environments. These algorithms provide additional means to determine person-centric metrics but do not address the team formation problem.

3 Approach

Expert team composition consists of three phases. First, we extract experts from discussion threads and form a social network (Fig. **??** Step 1). Basically, each reply to a posting results in an edge between the author of the original posting and the reply's author.

Next, we derive expert skill profiles from titles and tags of discussion threads (Fig. **??** Step 2). Each word represents a particular skill. Identification of meaningful words from tags and titles is considered outside the scope of this work as ultimately it is up to the user which skills (i.e., words) he/she defines as required to be provided by the team of experts. The number of postings in a thread associated with a particular skill determines the expert's skill level.

Finally, we solve the team formation subproblem applying a heuristic that searches the generated social network for the optimum team (Fig. **??** Step 3). The optimum team configuration depends on the trade-off between skill coverage and expert connectivity. We can recommend the top expert for each skill or accept less qualified team members which, however, yield a more tightly connected

interaction network. The motivating example in the following subsection outlines these steps in more detail.

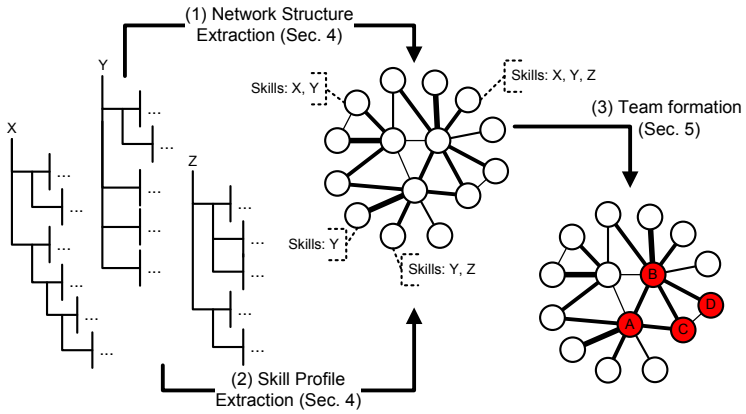

Fig. 1: Extracting network structure (1) and skill profile (2) from discussion threads for expert team formation (3).

Motivating Example We observe the postings of five experts (Alice, Bob, Carol, Dave, and Eve) across three posting threads on mobile computing (*Mobile*), web technologies (*Web*), and streaming mechanisms (*Streaming*). Fig. **??** (left) provides the posting structure of these three discussions, displaying only the author of a posting. The corresponding interaction structure is outlined in Fig. **??** (right). In the discussion on mobile technologies Alice replied to a posting by Bob. Therefore, we create an edge between these to experts. Carol also participated in the discussion but did not respond directly to any of the observed experts. Consequently we raise only her expertise level without creating any edges. The same holds true for Alice's second posting. Eventually we derive following top experts for the three skills *Mobile*, *Web*, and *Streaming*: Alice is the top expert for *Mobile* with two postings, Eve yields best expertise for *Web* with three postings, and Bob and Carol share the most postings for *Streaming*.

We select Alice, Carol, and Eve to obtain one potential best qualified team. They have, however, never directly interacted and thus share no common edge in the network. An immediate improvement results from exchanging Carol for Bob who is connected to Alice and yields the same expertise level as Carol. Ultimately, a sensible trade-off consists of introducing Dave to the team for providing the *Web* skill although he is not the most qualified expert. This trade-off, however, produces an even better connected team where everyone knows everyone else. In this paper, we will also consider edge weights to derive tighter connected teams. We ignore them here, as the edge weights are all equal in this example.

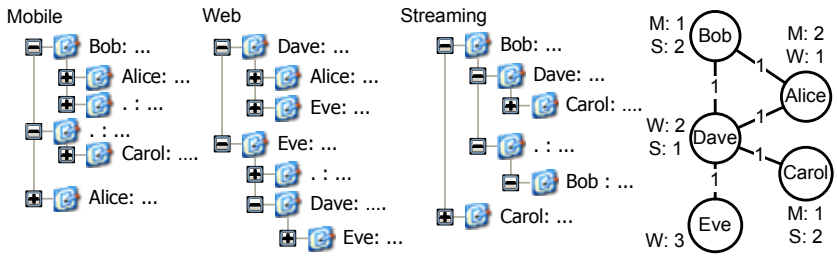

Fig. 2: Example transformation of discussion threads to interaction network and skills. (Edge labels equal interaction count; M/W/S: *Mobile*, *Web*, and *Streaming* skills.)

4 Network and Skill Extraction

A posting thread is mapped to a tree graph $\mathcal{F}(\mathcal{P}, \mathcal{R})$ with postings $p \in \mathcal{P}$ linked by directed, weighted edges $r \in \mathcal{R}$. Any edge r_i points from a reply to its parent posting and initially yields weight $w(r_i) = 1$.

The expert social network is modeled as an undirected graph $\mathcal{SN}(\mathcal{V}, \mathcal{E})$. The nodes are experts $v \in \mathcal{V}$, the weighted edges $e \in \mathcal{E}$ link two experts if at least one of them has once posted a reply to the other. Applying undirected edges is more appropriate than directed edges as without in-depth semantic analysis we cannot distinguish between an expert correcting a novice, or a novice requesting clarification from an expert. We, therefore, interpret each link as an interaction. The edge weight $w(e)$ derives from the amount of replies between two experts.

Each expert exhibits a skill profile $\mathcal{SP}$ comprising of multiple skills s. During thread transformation, we first extract the topics from the thread and interpret them as skills[3]. We then count for each expert active in the underlying thread the number of postings and increase his/her skill counter $k_i(s)$ correspondingly. Each posting increases the skill counter by 1 as this is the simplest assumption when the content of the posting remains unknown. Some online discussion sites apply a moderation scheme that allows moderators and/or other users to evaluate the quality of individual postings (e.g., `http://slashdot.org`). Filtering out of low quality posts can then be applied to better represent a user's expertise and deter non-experts to boost their expert-level through flooding a forum with irrelevant posts. Scores provide a refined expertise structure, however our general approach works on any discussion tree.

The absolute skill values are only an intermediary metric, as we need to be able to compare multiple skills. To this end, we measure for each expert and each skill the expertise level $q_i(s)$ in the interval $[0; 1]$. A linear transformation maps the absolute skill counter to the relative expertise level: $q_i(s) = k_i(s)/max(k(s))$ such that the expert with most postings of a given skill s yields expertise $q(s) = 1$. The overall aggregation of skills from every expert determines the

[3] Consideration of synonyms and/or additional skill reasoning based on knowledge models is outside the scope of this paper.

network's skill portfolio $\mathcal{S}_{\mathcal{SN}}$. Note that the transformation from thread structure to interaction network doesn't necessarily create a new social network but rather updates the edges between experts as well as skills of experts in an already existing network. Transformation of the posting structure to social network links is, however, not immediately applicable on raw threads in the presence of anonymous postings.

Thread Reduction Most discussion sites allow anonymous postings. The challenge is to remove those postings without jeopardizing the transformation of remaining postings to the interaction network. There are two ways to deal with such postings. On the one hand, we can ignore them and rely solely on direct replies between known experts. When the number of anonymous postings is high, however, this will reduce the likelihood of having a sensible set of interactions. On the other hand, we can simply bridge the anonymous postings, thereby risking to introduce interactions between experts that do not reflect reality. Take the initial thread in Figure **??** (left part) as an example. In the first case, we would merely derive a link between Eve and Dave. The second case would yield an overrated link between Eve and Dave (4 replies) and also an overly strong link between Dave and Alice (1 reply).

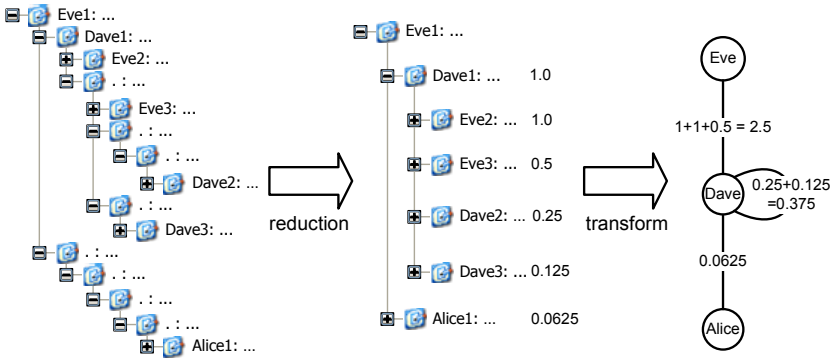

Fig. 3: Reduction of posting threads and transformation to interaction network.

Our thread reduction algorithm (Alg. **??**) mitigates both disadvantages by combining link bridging with link dampening. The link strength between a posting and a reply is halved for each intermediary anonymous posting, thereby rapidly decreasing with distance. Figure **??** (middle part) displays the reduced thread and the link strength of each reply towards its parent. Ultimately, all non-anonymous users from a discussion tree become part of the social network. When extracting an optimal team, however, filtering out experts below a minimum level of expertise improves processing speed.

Algorithm 1 Thread Reduction Algorithm $\mathcal{TRA}(\mathcal{F}(\mathcal{P}, \mathcal{R}))$.

```
for Posting m ∈ P do
    if isAnonymousCreator(m) then
        Posting parent ← getSuccessor(F, m)
        ReplyEdge mp ← getEdge(T, m, parent)
        /* Remove the edge from the anonymous posting to its parents */
        removeEdge(F, mp)
        for Posting child ∈ getPredecessors(F, m) do
            /* For each reply */
            ReplyEdge cm ← getEdge(F, child, m)
            /* Remove the edge from the reply to the posting */
            removeEdge(F, cm)
            /* Add a new reply edge bridging the anonymous posting */
            ReplyEdge cp ← createEdge(F, child, parent)
            /* Reduce the edge weight */
            setEdgeWeight(cp, getWeight(mp) * getWeight(cm) * 0.5)
        end for
    end if
end for
```

5 Formalizing the Team Formation Problem

The team formation problem defines a set of required skills $\mathcal{S}_\mathcal{R} \subseteq \mathcal{S}_{\mathcal{SN}}$. The importance of each contained skill $s \in \mathcal{S}_\mathcal{R}$ is given by weight $w(s)$ in the interval $]0;1]$ with $\sum_i w_i(s) = 1$. The goal is finding the set of experts $\mathcal{T} \subseteq \mathcal{V}$ that exhibits a good-enough match of the required skill while providing a sufficient degree of connectivity within the team. Here the exact meaning of the terms *good-enough* and *sufficient* are subject to the trade-off between *skill coverage* and *team distance*.

Skill coverage measures how well the set of experts match the required set of skills $\mathcal{S}_\mathcal{R}$. For each skill s, the best match is the team member with the highest corresponding expertise level $q(s)$. A single expert potentially yields the highest expertise level for multiple skills. Subsequently, the team's overall skill fulfillment $\mathcal{C}_\mathcal{T}$ is defined as:

$$\mathcal{C}_\mathcal{T} = \sum_i max(q_j(s_i)) * w(s_i) \qquad where\ v_j \in \mathcal{T} \qquad \forall s_i \in \mathcal{S}_\mathcal{R} \tag{1}$$

The team with maximum achievable coverage is denoted as $Top(\mathcal{S}_\mathcal{R})$ and yields for every required skill an expert with maximum expertise level $q(s) = 1$. The top team, however, is not the best choice when its members have little interacted before. The team distance $\mathcal{D}_\mathcal{T}$ quantifies the amount and strength of inter-team links compared to the maximum possible connectivity:

$$\mathcal{D}_\mathcal{T} = \sum_{\mathcal{E}_\mathcal{T}} \frac{1}{w(e_{ij})} + (\frac{|\mathcal{T}| * (|\mathcal{T}| - 1)}{2} - |\mathcal{E}_\mathcal{T}|) * \beta * max(\frac{1}{w(e)}) \qquad \forall i, j \in \mathcal{T} \tag{2}$$

where $|\mathcal{E}_\mathcal{T}|$ is the number of intra-team edges whereas $max(\frac{1}{w(e)})$ determines the weight of the weakest link in the overall social network. As we aim to minimize $\mathcal{D}_\mathcal{T}$, we sum across all inverted edge weights and add a penalty for every non-existent edge. The penalty parameter β determines the impact of such a non-existing edge. For $\beta = 1$ a missing edge between two experts receives the minimum interaction weight, thus switching to a configuration with an additional edge has almost no impact on distance. As $\beta \to \infty$ the existing edge weights loose their significance and teams exhibiting a fully connected graph will yield the lowest distance. A sensible value derived from our experiments is $\beta = 4$ which we will use throughout this paper.

The overall team quality $\mathcal{Q}_\mathcal{T}$ is ultimately obtained though aggregating skill coverage $\mathcal{C}_\mathcal{T}$ and team distance $\mathcal{D}_\mathcal{T}$. We introduce the trade-off parameter α that configures acceptable combinations of coverage and distance:

$$\mathcal{Q}_\mathcal{T} = \alpha * \mathcal{C}_\mathcal{T} + (1 - \alpha) * (1 - \frac{\mathcal{D}_\mathcal{T}}{\mathcal{D}_{MAX}}) \qquad \alpha = [0; 1] \tag{3}$$

where $\mathcal{D}_{MAX}$ is the maximum distance for a team of $|\mathcal{S}_\mathcal{R}|$ experts what yield no direct inter-team edges ($|\mathcal{E}_\mathcal{T}| = 0$). The top team $Top(\mathcal{S}_\mathcal{R})$ will yield the highest quality when α approaches 1, whereas the best connected team will provide the best quality for $\alpha \to 0$. We can guarantee a minimum skill coverage level if we include only experts that exhibit a given expertise threshold.

For the example team formation problem in Figure **??**, we derive following quality measurements for the three considered teams when applying $\alpha = 0.5$. A team comprising Alice, Carol, and Eve will yield quality $\mathcal{Q}_{ACE} = 0.5$ with $\mathcal{C}_{ACE} = 1$ and $\mathcal{D}_{ACE} = 12$. The team consisting of Alice, Bob, and Eve achieves better distance, and thus also better quality: $\mathcal{Q}_{ABE} = 0.625$, with $\mathcal{C}_{ABE} = 1$ and $\mathcal{D}_{ABE} = 9$. Finally, the combination of Alice, Bob, and Dave provides the best quality (for the given trade-off parameter): $\mathcal{Q}_{ABD} = 0.82$, with $\mathcal{C}_{ABD} = 0.\mathring{8}$ and $\mathcal{D}_{ABD} = 3$. In all three cases, $\mathcal{D}_{MAX} = 12$ as 3 *edges* $*$ $(\beta = 4)$ $*$ $(max(1/w(e)) = 1) \to 12$.

6 Team Formation Heuristic

In the search of a better trade-off between skill coverage and team distance, we need to test various expert combinations. Investigations of the rich-club phenomenon in scientific collaboration networks (e.g., [**?**]) have shown that a sufficiently well connected team is unlikely to be amongst the very top ranked experts. A network exhibiting rich-club properties has the best-connected nodes form tightly connected communities. In the case of expert networks — such as scientific author networks — such tight collaborative groups exist only within particular research domains but not beyond. Consequently, we need to include also experts below the top 10 in our search for acceptable team configurations when skills are increasingly different (e.g., when skills belong to distinct domains).

Brute-force testing of every possible combination, however, quickly becomes unfeasible. Testing the top m experts for $\mathcal{S}_\mathcal{R}$ skills has $\mathcal{O}(m^{\mathcal{S}_\mathcal{R}})$ computational complexity (i.e., already for 10 experts and 10 skills, we would need to analyze 10 billion combinations). Our goal is to find a better connected team than the aggregation of the top experts for each skill but not necessarily the best possible solution. Simulated Annealing [**?**] is a suitable optimization heuristic for this problem.

6.1 Simulated Annealing

Simulated Annealing (SA) is a heuristic for approximating a global optimum in complex mathematical problems. It is well suited for problems with discrete search space such as the order of cities in the traveling sales man problem. We briefly outline the generic heuristic aspect and discuss the problem specific parts in the subsequent subsections in more detail. Simulated annealing is an iterative process building on following basic components:

Candidate Solution contains the current best problem solution which is gradually improved. In the team formation problem, the current solution $\mathcal{T}$ assigns one expert to each skill, potentially having one expert covering multiple skills.

Solution Energy Function measures the quality of a given solution. SA aims to find a solution with the lowest possible energy. The current quality function $\mathcal{Q}_\mathcal{T}$ returns higher values for better team configurations. We provide the restructured equation in the following subsection to derive a suitable energy function.

Neighborhood Function provides a new candidate solution based on the current solution. A good neighborhood function traverses the search space quickly, but produces new solutions that yield similar energy level to the preceding solution. The neighborhood function takes a team configuration and replaces the expert of a random skill.

Transition Function decides whether to accept a new solution or to stick with the current one. Simulated Annealing also accepts team configurations that yield worse quality than the current one to avoid local optima.

Cooling function gradually reduces the temperature. Large solution changes are less likely for lower temperatures. As the temperature falls, worse solutions are less likely to be accepted.

We briefly outline the iterative process in Algorithm **??** as provided in the JUNG 1.7.6 framework[4]. We omit some configuration parameters for sake of clarity. Transition function and Cooling function are problem independent, thus introduced here. We discuss neighborhood function and energy function in the subsequent subsections. For now, we treat these as blackboxes.

Simulated annealing takes an initial solution (i.e., the top expert for every skill) and derives the corresponding energy. Simulated Annealing continues to

[4] `http://jung.sourceforge.net/`

evaluated similar solutions as long as the temperature has not reached zero and there are more available iterations. A new solution is always accepted when it yields lower energy (Alg. **??** line 12). Worse solutions are accepted with probability p_{SA} defined as:

$$p_{SA} = e^{-\frac{\delta_{energy}}{temp}} \tag{4}$$

where δ_{energy} is the energy difference between the current and new solution, $temp$ is the current annealing temperature, and e is Euler's number. A transition to a solution with higher energy is possible as long as the temperature remains high, or the energy difference is very small.

The freezing process depends on the cooling rate and current iteration state. As long as the number of successful transitions is high (i.e., *success* close to *tries*) the system remains in a search space region that still provides many solutions with lower energy. The function for the temperature in the next iteration is defined as:

$$temp_n = r_{cooling}^{(limit_{accept} - \frac{success}{tries}) * tries} * temp \tag{5}$$

where $tries$, $r_{cooling}$, and $limit_{accept}$ are configuration parameters. For our experiments, we apply $tries = 100$, $r_{cooling} = 0.99$, and $limit_{accept} = 0.97$

Algorithm 2 Simulated Annealing Algorithm $\mathcal{SA}(maxIt, startTemp)$.

```
1: T ← calcNewSolution(startTemp)
2: nrg ← calcEnergy(T)
3: temp ← startTemp
4: iteration ← 0
5: while temp > 0 ∪ iteration < maxIt do
6:     success ← 0
7:     for tries do
8:         /* Neighborhood function provides a new solution. */
9:         newSolution ← calcNewSolution(T, temp)
10:        nrg_new ← calcEnergy(newSolution)
11:        δ_energy ← nrg − nrg_new
12:        if doTransition(δ_energy, newSolution, temp) then
13:            T ← newSolution
14:            nrg ← nrg_new
15:            success + +
16:        end if
17:    end for
18:    temp ← calcTemperature(temp, success)
19:    iteration + +
20: end while
21: return T
```

6.2 Simulated Annealing Energy Function

The energy function provides the tradeoff between skill coverage $\mathcal{C}_\mathcal{T}$ and team distance $\mathcal{D}_\mathcal{T}$. A solution consists of an expert for each skill. We cannot directly reuse the overall team quality function $\mathcal{Q}_\mathcal{T}$ as SA requires an energy function that decreases with raising solution quality. The initial solution consists of the top expert for each required skill in $\mathcal{S}_\mathcal{R}$. This composition provides an upper boundary to the possible skill coverage. Any better solution must exhibit lower energy by reducing the distance $\mathcal{D}_\mathcal{T}$. Expert compositions that additionally come with lower coverage need to yield proportionally even lower distance. The proportion is determined by the tradeoff factor α. The initial solution has energy = 1. Any solution that reduces coverage and distance to similar extent yields also energy = 1. Ultimately, the energy function for a solution $\mathcal{T}$ is defined as:

$$nrg_\mathcal{T} = \frac{1 - (\alpha * \mathcal{C}_\mathcal{T} + (1-\alpha) * (1 - \mathcal{D}'_\mathcal{T} * \mathcal{D}_{top}^{-1}))}{1-\alpha} \tag{6}$$

where $\mathcal{D}'_\mathcal{T}$ is the normalized distance $\mathcal{D}_\mathcal{T}/\mathcal{D}_{MAX}$ and $\mathcal{D}_{top}$ is the normalized distance of the initial solution. As no solution can yield higher coverage than the top experts, any solution with higher distance than $\mathcal{D}_{top}$ will yield an energy value greater than 1 and thus can safely be ignored. Dividing the aggregation of skill coverage and distance by $1-\alpha$ ensures that regardless of α the initial solution and any proportional tradeoff will always yield $nrg = 1$.

6.3 Simulated Annealing Neighborhood Function

The neighborhood function generates a new solution given a current solution. The function needs to be able to (a) traverse the search space in short time and (b) find neighboring configuration with similar energy. The first requirement guarantees that the simulated annealing algorithms is able to reach all states in a timely manner, thus potentially identifying the optimum solution. The second requirement ensures the algorithm's convergence. A random solution is more likely to be worse (rather than better) than the current solution. Jumping between high energy states maintains a high temperature level, thereby keeping the system from cooling down and finding the desired areas of low energy.

Our neighborhood function addresses both concerns. We randomly select a required skill s and exchange the current expert v_{old} with another expert v_{new} with probability p_{nh}. The neighborhood probability p_{nh} depends on the interaction proximity and threshold parameter s_{nh}

$$s_{nh} = \frac{temp}{maxTemp} * (prox_{max} - prox_{min}) + prox_{min} \tag{7}$$

$$p_{nh}(e_{new}) = \begin{cases} \frac{1}{m-1} & if\ prox(v_{old}, v_{new}) \geq s_{nh} \\ \frac{\psi}{m-1} & otherwise \qquad with\ \psi = \frac{prox(v_{new}, v_{old}) - prox_{min}}{s_{nh} - prox_{min}} \end{cases} \tag{8}$$

where m is the number of candidate experts. The proximity $prox(v_{old}, v_{new})$ between two experts is defined by the shortest hop path (SHP) with maximum

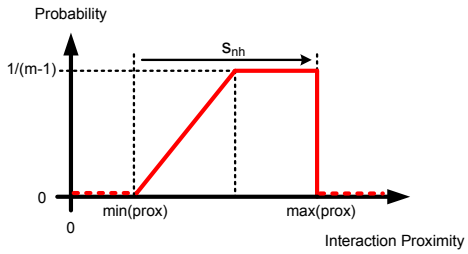

Fig. 4: Probability function for selecting a particular expert based on annealing temperature and interaction proximity.

edge weights. We sum across all traversed edges and take the hop count into account to penalize for the number of intermediary experts. The stronger two experts are connected, the higher their proximity.

$$prox(v_i, v_j) = \frac{\sum_k w(e)}{k * (k-1) + 1} \qquad \forall w(e) \in max[SHP(v_i, v_j)] \tag{9}$$

For the expert selection probability p_{nh} we define $prox_{max} = max[prox(v_{old}, v_i)]$ and analog $prox_{min} = min[prox(v_{old}, v_i)]$ where v_i is any expert that provides a minimum expertise level of the selected skill ($q(s) > 0$). This prevents the selection of experts that are in close proximity but who do not provide the required skill. When the expected workload requires a minimum number of members t_{min}, we simply remove members of the current solution from the candidate set. We first remove the worst ranked existing expert until no expert from the candidate set (if selected) would violate the size constraint.

The neighborhood probability function ensures that experts that are in proximity of the current solution are more likely to be selected, than experts further away. Besides interaction proximity, also the current temperature affects this probability. In the beginning, when temperature is still high, proximity has little effect and every expert is equally like being selected. Later in the process, the probability of selecting a particular expert decreases linearly with distance. As shown in Figure **??**, s_{nh} moves from $prox_{min}$ to $prox_{max}$ as the temperature falls towards zero. Note that the neighborhood probability function p_{nh} is not a classical probability density function as the sum of probabilities for all observed experts does not add up to 1.

This neighborhood function enables to quickly traverse the complete search space at the beginning. Later, we still can reach every solution, but require more steps to do so. We assume two experts in proximity to yield similar links to common neighbors. Thus, as we increasingly select new experts that are close to their replaced predecessor, the total connectivity will improve on average more than selecting random experts. Subsequently, two candidate solutions will yield similar energy values. This avoids fruitless testing of solutions with high energy.

7 Evaluation

We evaluate our team formation mechanism with a real world data set extracted from Slashdot. We provide a brief introduction to the data set. The experiments consist of 8 sets of 5 skill configurations for a total of 45 different skill configurations. We present one example for in-depth discussion of the effect and successful results of our approach.

Experiment Setup Slashdot is a well understood and rich data set [**?**] describing a large user community. Users submit information technology related news items which the editors decide to publish or not. News fall into multiple categories (i.e., subdomains) such as *linux*, *apple*, or *games*. A published piece of news becomes a *story* which all users—anonymous or logged-in—can comment on. These comments create a posting hierarchy. Slashdot exhibits the characteristics of a large-scale expert network. Some users remain consistently active throughout all subdomains. Other users join in an ad-hoc manner, participate for a limited period, and then vanish again. Users are interested in providing their knowledge to improve the quality and information content of a story. They rarely engaging in long running personal communication threads with other users [**?**,**?**].

The subdomain names are directly mapped to skills. Each story maps to the skill represented by its parent subdomain. The slashdot moderation system also enables classification of postings according to *Insightful*, *Interesting*, *Informative*, *Funny*, etc. content which we combine with subdomains to generate more fine-grained skills. We group the experiment set in two rough categories: (i) cross-subdomain teams and (ii) mixed inter-intra subdomain teams.

The first three experiment sets ($Ex_1 \rightarrow Ex_3$) yield 6, 7, and 8 skills, respectively, out of 10 available subdomains. When ever a user posts in a story within a subdomain, his/her corresponding skill is raised by 1. The first skill set $\mathcal{S}_{SUB}$ contains $S1 = apple$, $S2 = ask$, $S3 = entertainment$, $S4 = mobile$, $S5 = linux$, $S6 = developers$, $S7 = games$, $S8 = news$, $S9 = slashdot$, and $S10 = it$. The remaining 5 experiment sets introduce skills combined from predicates and subdomains; thus only postings with predicates are considered. When a user's posting within subdomain X receives a predicate Y, then the user's skill XY is increased (e.g., an *Interesting* posting within subdomain *linux* increases skill *linuxInteresting*. The experiment sets ($Ex_4 \rightarrow Ex_8$) are derived from 2x4, 6x2, 4x3, 3x4, and 4x4 subdomain-predicate combinations. The second skill set is the combination of $\mathcal{S}_{SUB}$ and $\mathcal{S}_{PRED} = \{P1 = Informative,\ P2 = Interesting,\ P3 = Funny,\ P4 = Insightful\ \} \rightarrow \{S1P1,\ \ldots,\ S10P4\}$ for a total of 40 skills.

All experiment rounds applied $\alpha = 0.5$ and experts had to yield $q(s) \geq 0.5$ to be considered for a particular skill s. Experts that did not provide a single required skills were temporarily removed from the social network to improve processing speed. Each skill was considered of equal importance, thus the skill weights $w(s)$ were uniformly set to $1/|S_R|$.

Experiment Results We analyze experiment set Ex_1 configuration 4 in more detail. The required eight skills are $S_R = \{S1, S2, S3, S4, S7, S8, S9, S10\}$. Figure **??** visualizes the interaction structure of the initial and the optimal team. The initial team of top experts are $T_{top} = \{V5, V6, V7, V8, V9, V10\}$. The optimal team $T_{opt} = \{V1, V2, V3, V4, V7, V10\}$ keeps $V7$ and $V10$ from the initial team but introduces four new experts[5]. Note that expert $V7$ and $V10$ provide two skills each (see also Table **??**). The optimal team improves the distance by 90% (from $D_{TOP} = 0.40$ down to $D'_T = 0.04$) while reducing the skill coverage to $C_T = 0.67 (\equiv 33\%)$. The optimal team graph is fully connected, and in addition, also exhibits stronger links between members than the initial team.

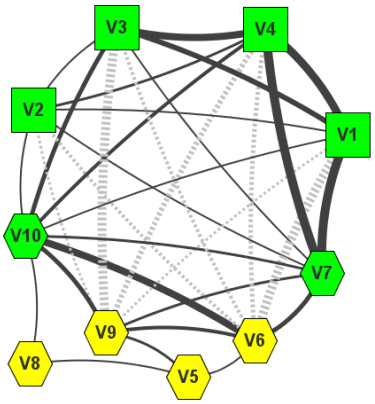

Fig. 5: Optimal expert team (green squares) and initial expert team (yellow hexagons) for experiment set Ex_1 skill configuration 4. Line thickness represents interaction distance; with solid intra-team links and dashed inter-team links. (Colors online)

Skill	T_{top}	T_{opt}	$r(s)$	$q(s)$
$S1$	$V8$	$V1$	5	0.594
$S2$	$V7$	$V7$	1	1.0
$S3$	$V6$	$V10$	12	0.535
$S4$	$V5$	$V10$	6	0.515
$S7$	$V9$	$V2$	4	0.524
$S8$	$V7$	$V7$	1	1.0
$S9$	$V10$	$V3$	5	0.577
$S10$	$V7$	$V4$	7	0.648

Table 1: Expert skills for initial T_{TOP} and optimal team T_{OPT} for Ex_1 configuration 4. Rank $r(s)$ and utility $q(s)$ of each expert for the provided skill are in brackets. All experts in T_{TOP} yield rank $r = 1$ and utility $q(.) = 1$.

We have printed the energy of the initial, optimal, and next top-30 solutions in Figure **??**. The best and second best solution yield almost identical energy and are also distance and skill fulfillment wise rather similar. The team configuration, however, differs by one dropped, and two additional members. Most top-30 solutions for this skill configuration remain around $nrg_T = 0.83$. Within those solutions a broad spectrum of teams exist that cover the range from low distance/low skill coverage to high distance/high skill coverage. We can then apply a-posteriori preferences towards higher skilled or lower skilled expert compositions from such a set of equally qualified teams. Our algorithm found in total

[5] These 10 experts make up T_{top} and T_{opt}. The corresponding slashdot IDs are: $V1$:622222, $V2$:987471, $V3$:595695, $V4$:71849, $V5$:25149, $V6$:912633, $V7$:957197, $V8$:18001, $V9$:727027, and $V10$:835522. In total, 49 experts qualified as team member candidate (at least one $q(s) > 0.5$) embedded in a social network of 17765 experts.

more than 100 team configurations which yield a better trade-off than the initial team T_{top}.

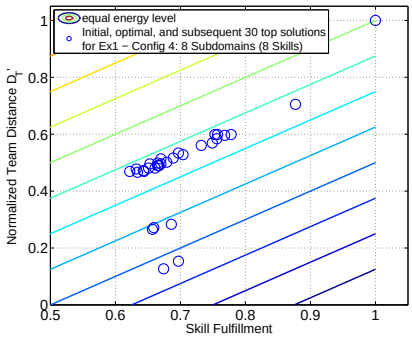

Fig. 6: Initial (top right), optimal (bottom left), and top 30 solutions (normalized values) for Ex_1 configuration 4.

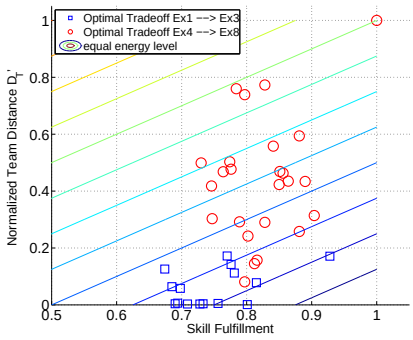

Fig. 7: Normalized optimal solutions for $Ex_1 \rightarrow Ex_3$ (squares) and $Ex_4 \rightarrow Ex_8$ (circles). (All initial solutions reside at $[1, 1]$.

The results of configurations $Ex_1 \rightarrow Ex_3$ reveal that the team formation heuristic finds significantly better solutions for all of the observed skill combinations. The normalized team distance D'_T of the optimal solutions (Figure **??** blue squares) amounts to 0.004 up to 0.17 of the initial distance while skill fulfillment remains comparatively high between 0.67 and 0.93. For each required skill configuration, our approach found more than 100 solution that all provide a better tradeoff than the initial team configuration T_{top}. When we compare the absolute distance values (D_T) in Figure **??**, we note that the initial distance does not affect the heuristic's ability to find significantly well connected teams.

In experiment sets $Ex_4 \rightarrow Ex_8$, we tested for the effect of correlated skills (i.e., predicates within the same subdomain) and larger skill sets. The heuristic still determines better solutions (except for configuration 5 in set 8 having 16 skills) but does not achieve as high quality tradeoffs as in $Ex_1 \rightarrow Ex_3$ (compare the red circles to the blue squares in Figure **??**). Increasing the skill set results in larger teams which are unlikely to exhibit similar dense connectivity as smaller groups. Larger skill sets come also with lower alternative trade-off solutions. For the $Ex_4 \rightarrow Ex_8$, there exist on average 84.2, 76.4, 48.6, 49.6, and 17.2 alternatives, respectively. The effect of similar skills becomes apparent in Ex_8 which contains all four predicates from 2 subdomains (ocher, upside down triangles in Figure **??**). Although team distance is significantly lower, the improvement remains en par with the worst improvements in experiment 1. The average distance of set Ex_8 $\overline{D'_{EX8}} = 0.23$ remains significantly above the distance of set Ex_1 $\overline{D'_{EX1}} = 0.08$ (having also 8 skills).

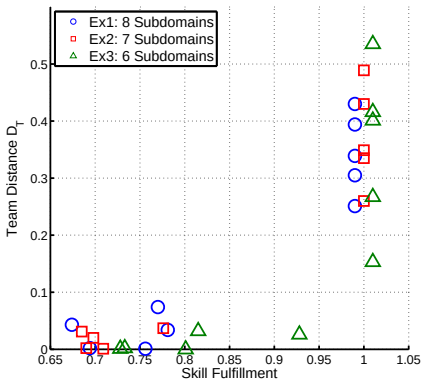

Fig. 8: Initial and optimal solutions for the 15 skill configurations in $Ex_1 \rightarrow Ex_3$. Initial solutions are slightly shifted (+/- 0.01) for sake of clarity.

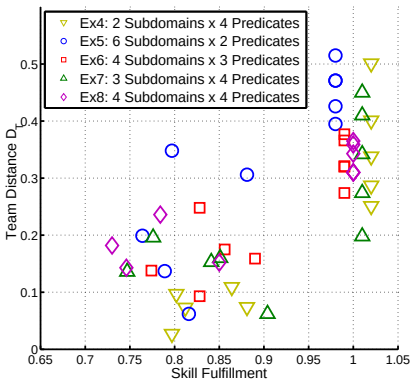

Fig. 9: Initial and optimal solutions for the 25 skill configurations in $Ex_4 \rightarrow Ex_8$. Initial solutions are slightly shifted (+/- 0.02) for sake of clarity.

Evaluation Summary Our team composition algorithm provides excellent results across all experiment sets. Figure **??** compares the average and standard deviation of energy, distance, and coverage for all eight experiment sets. Our approach yields for all set consistently high skill coverage ($\sim$ 0.8). The overall energy is, hence, mostly determined by team distance. Increasing skills and skill similarity limit the achievable distance reduction. We are able to detect significantly better connected teams for configurations of up to 12 skills, and still some improvement for up to 16 skills. Teams without explicit management structure, however, rarely exceed 10 to 12 members. We therefore did not test any configurations larger than 16 skills.

8 Conclusion

Online communities provide both the raw data for deriving expert skill profiles and their interaction structure. Considering only independent expert properties is insufficient for finding the best suited team. Optimal team composition requires a trade-off between skill coverage and expert connectivity. We have demonstrated the benefit of our heuristic for finding well connected experts that simultaneously yield high expertise level in a social network. In the future, we will investigate mechanisms to support skill dependencies (i.e., respective experts connected tighter). This allows for finding optimally structured teams without having to focus on neither fully connected graphs nor minimum-spanning tree graphs. Such skill dependencies also provide the basis for composing optimally structured teams beyond 12 members in combination with management skills. At the same time, we plan to evaluate our algorithm with other online commu-

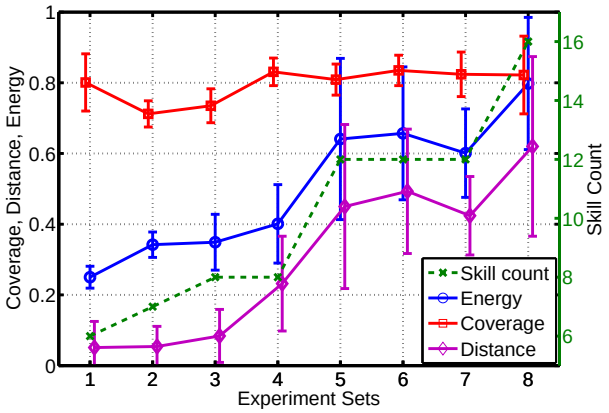

Fig. 10: Average and standard deviation of energy, distance, and coverage for all experiment sets (full lines). The respective skill count for each experiment set is given by the dashed, green line. (Colors online)

nities that exhibit a larger overall skill set. A qualitative comparison to other team formation algorithms is also an open topic.

Acknowledgment The authors would like to thank Daniel Schall and Florian Skopik for providing the slashdot data set. This work has been partially supported by the EU STREP project Commius (FP7-213876).

References

1. Adamic, L.A., Zhang, J., Bakshy, E., Ackerman, M.S.: Knowledge sharing and yahoo answers: everyone knows something. In: WWW '08: Proceeding of the 17th Int. Conference on World Wide Web. pp. 665–674. ACM, New York, NY, USA (2008)
2. Backstrom, L., Huttenlocher, D., Kleinberg, J., Lan, X.: Group formation in large social networks: membership, growth, and evolution. In: KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 44–54. ACM, New York, NY, USA (2006)
3. Baresi, L., Bianchini, D., Antonellis, V.D., Fugini, M.G., Pernici, B., Plebani, P.: Context-aware composition of e-services. In: TES. pp. 28–41 (September 2003)
4. Baykasoglu, A., Dereli, T., Das, S.: Project team selection using fuzzy optimization approach. Cybern. Syst. 38(2), 155–185 (2007)
5. Bird, C., Gourley, A., Devanbu, P., Gertz, M., Swaminathan, A.: Mining email social networks. In: MSR '06: Proceedings of the 2006 Int. Workshop on Mining software repositories. pp. 137–143. ACM Press, New York, NY, USA (2006)
6. Bird, C., Pattison, D., D'Souza, R., Filkov, V., Devanbu, P.: Latent social structure in open source projects. In: SIGSOFT '08/FSE-16: Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering. pp. 24–35. ACM, New York, NY, USA (2008)

7. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems 30(1-7), 107–117 (1998), proceedings of the Seventh International World Wide Web Conference
8. Cheatham, M., Cleereman, K.: Application of social network analysis to collaborative team formation. In: CTS '06: Proceedings of the International Symposium on Collaborative Technologies and Systems. pp. 306–311. IEEE Computer Society, Washington, DC, USA (2006)
9. Colizza, V., Flammini, A., Serrano, M.A., Vespignani, A.: Detecting rich-club ordering in complex networks. Nature Physics 2, 110–115 (2006)
10. Dustdar, S., Schreiner, W.: A survey on web services composition. Int. J. Web Grid Serv. 1(1), 1–30 (2005)
11. Fitzpatrick, E.L., Askin, R.G.: Forming effective worker teams with multi-functional skill requirements. Comput. Ind. Eng. 48(3), 593–608 (2005)
12. Gaston, M.E., Simmons, J., desJardins, M.: Adapting network structure for efficient team formation. In: AAMAS-04 Workshop on Learning and Evolution in Agent Based Systems (July 2004)
13. Gómez, V., Kaltenbrunner, A., López, V.: Statistical analysis of the social network and discussion threads in slashdot. In: WWW '08: Proceedings of the 17th Int. Conference on World Wide Web. pp. 645–654. ACM, New York, NY, USA (2008)
14. Haveliwala, T.: Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search. IEEE Transactions on Knowledge and Data Engineering 15(4), 784–796 (July-Aug 2003)
15. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science, Number 4598, 13 May 1983 220, 4598, 671–680 (1983)
16. Lappas, T., Liu, K., Terzi, E.: Finding a team of experts in social networks. In: KDD '09: Proceedings of the 15th ACM SIGKDD Int. Conference on Knowledge discovery and data mining. pp. 467–476. ACM, New York, NY, USA (2009)
17. Maamar, Z., Benslimane, D., Thiran, P., Ghedira, C., Dustdar, S., Sattanathan, S.: Towards a context-based multi-type policy approach for web services composition. Data Knowl. Eng. 62(2), 327–351 (2007)
18. McAuley, J.J., da Fontoura Costa, L., Caetano, T.S.: Rich-club phenomenon across complex network hierarchies. Applied Physics Letters 91(8), 084103 (2007)
19. Quitadamo, R., Zambonelli, F., Cabri, G.: The service ecosystem: Dynamic self-aggregation of pervasive communication services. In: Software Engineering for Pervasive Computing Applications, Systems, and Environments, 2007. SEPCASE '07. First International Workshop on. pp. 1–10 (May 2007)
20. Schall, D.: Human Interactions in Mixed Systems - Architecture, Protocols, and Algorithms. PhD Thesis, Vienna University of Technology, Karlsplatz 13, 1040 Vienna, Austria (2009)
21. Skopik, F., Schall, D., Dustdar, S.: Modeling and mining of dynamic trust in complex service-oriented systems. Information Systems Journal 35(7), 735–757 (10 2010)
22. Skopik, F., Truong, H.L., Dustdar, S.: Trust and reputation mining in professional virtual communities. In: 9th International Conference on Web Engineering (ICWE). Springer (June 2009)
23. Wi, H., Oh, S., Mun, J., Jung, M.: A team formation model based on knowledge and collaboration. Expert Syst. Appl. 36(5), 9121–9134 (2009)
24. Yang, Y., Mahon, F., Williams, M.H., Pfeifer, T.: Context-aware dynamic personalised service re-composition in a pervasive service environment. In: UIC. pp. 724–735 (2006)