Crowdsourcing Recommendations From Social Sentiment

Yusheng Xie, Yu Cheng, Daniel Honbo, Kunpeng Zhang, Ankit Agrawal, Alok Choudhary
Department of Electric Engineering & Computer Science
Northwestern University
{yxi389,ych133,dkh301,kzh980,ankitag,choudhar}@eecs.northwestern.edu
2145 Sheridan Road, Evanston, IL 60201 USA

ABSTRACT

In this paper, we investigate an innovative recommendation system by incorporating relevant social opinion and sentiment information. Our recommendation system, a powerful application of social sentiment analysis, differs from many existing models, which investigate the situation where the social network itself is structured to work with the product ranking and is specially built inside an e-commerce website. In contrast, our proposed system focuses on constructing and inferring product recommendations from external social network services (SNS) such as Facebook. In our system, we process product features in a finite-dimensional polynomial linear space. Additional components of our proposed system include an asymmetric similarity measurement and an asymmetric advantage measurement. We also show that our definitions for the two measurements include specific properties that reduce the computational overhead in the experiments. An important aspect of our modeling is to incorporate user-generated high-level semantic sentiment in the analysis. We apply our models to real time data and observe promising results for not only product recommendation but also job recommendation.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering

General Terms Algorithms, Performance

Keywords Business Intelligence, Crowdsourcing

1. INTRODUCTION

Thanks to the growing availability of the Internet, social media websites and e-commerce websites have gained remarkable growth in recent years. Increasing numbers of goods are being sold on the Internet both in terms of quantity and categories. More importantly, increasing e-commerce activities induce an unprecedented amount of comments and reviews regarding their shopping experience and the products. In terms of user-generated contents, SNSs like Facebook are usually very popular and have more information than a built-in review and rating network (RRN) such as Amazon.com's reviewer community, although RRNs are often much better structured for data mining. In order to make a purchase decision among millions of products and relevant comments, online customers rely heavily on the results returned by automatic search engines. Traditional product recommendation

processing depends heavily on its built-in RRN without taking advantage of any external SNSs, which can be problematic for some e-commerce business. For example, unlike Amazon.com, BestBuy suffers from its not-so-active built-in RRN. It would be advantageous if BestBuy can somehow utilize the existing vast network of Facebook, where most of its retailing brands (e.g. Samsung, Dell, etc.) maintain high profile social engagement with the Facebook users. BestBuy could then capitalize from the large and active social services provided by Facebook without advertising or even deploying its less popular built-in RRN at all.

It is challenging to use opinion and sentiment information from a semi-structured dataset like the Facebook public fan pages (commonly known as Facebook Walls) in item ranking. Data mining from semi-structured SNS is in general a challenging task. One cannot simply apply sentiment algorithms to the text in SNS and expect miracles to happen. In fact, finding a sensible way to redirect the power of sentiment analysis to help understand the SNS generating the sentiment as a whole is highly non-trivial and valuable. Comparing the structures of a typical built-in RRN and an external SNS in Table 1 makes it clear why the latter is a much more challenging situation. Unlike RRN, most information from SNS requires further treatment in order to gain any valuable insights. In many cases, such treatment requires arduous work.

Table 1. Comparison between RRN and SNS

Feature	RRN	SNS
Reviewer opinion	Numeric rating	High-level sentiment
Pros vs. Cons	Labeled	Unlabeled
Review text	Relevant	Comment not always relevant
Reviewer information	Partial	None

Furthermore, each SNS is organized in different formats depending on its source. For example, Twitter and Facebook services allow their users to socialize in different ways. Without loss of generality and for better clarity, we restrict the SNS in our model to be Facebook-style public walls. SNSs with Facebook-style format are very popular in practice and provide enough data for our experiments. For example, Google+ and LinkedIn both adopt this style in their SNSs to allow companies or organizations to set up public pages. In general, a public SNS is a collection of public walls set up by companies or organizations. Each of these walls contains official posts; and each post can attract users to comment, like/unlike others' comments. As shown in Figure 1, Amazon Kindle's public wall contains multiple official posts; Facebook users can leave comments on each of the posts, which are circled in red. In Figure 2, we show examples of public comments made by Facebook users in response to a public post;
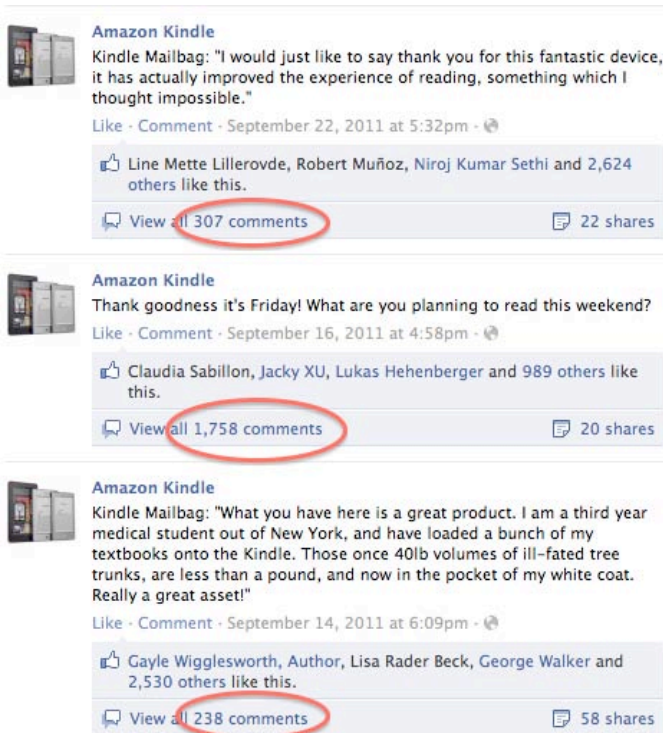
Figure 1. Snapshot of a typical Facebook public wall.

Figure 2. Snapshot of typical user comments to a Facebook post.

received by some comments from other users are circled in blue.

For simplicity, we further assume that an actual product for sale on an e-commerce website only corresponds to one Facebook public wall. For example, the product Amazon Kindle Fire only corresponds to the Facebook Amazon Kindle Fire wall and does not correspond to the Facebook Amazon Kindle wall.

2. RELATED WORK
The close relationship of our work with existing literature can be divided into two sections: its data and its model.

2.1 Data From Social Networks
The field of recommendation systems never lacks novel approaches especially when it is intertwined with opinion and sentiment information from SNS. The usage of social sentiment in this paper is largely inspired by the abundance of existing literature. The central idea in [14] and [15] produces recommendations from products on Amazon.com by extracting features from online product reviews. This work is particularly important to ours and is in fact employed in our model because it proposes a systematic way of mining high-level sentiment and semantic information in online text such as product reviews, tweets, etc. Later, [18] further refine the techniques for extracting high-level sentiment information by incorporating domain-specific prior knowledge or constraints in the context of social media. Understanding what people are saying on the Web and having a systematic way of categorizing the meaning are critical in our construction of recommendations. There are also a handful of high quality publications on the diversity of product recommendations, whose evaluation results we use in our experiments [19]. Balancing diversity and relevance can be a reasonable, although not irrefutable, metric for measuring the performance of a recommendation system particularly when the relevant system is not suitable for traditional measurements. Relating SNS to product recommendation is a hot topic in recent publications. [4], [8], [9], [16] and [13] study this topic from

interesting angles. The majority of such publications develop different models assuming a graph-type social network and most of them seem to understand and, in many cases, address the computational issues with large social networks. We think deploying social sentiment to play a central role in producing recommendations is still rarely explored and therefore entails our work. Last but not least, community formation and detection is also closely related to the field of SNS interacting with recommendation systems ([10], [11], and [23]).

2.2 Related Models
Perhaps the most frequent and the most famous word in recommendation system is "Collaborative Filtering (CF)". The model in our paper does not directly apply CF on any data, but the root of our model can be traced back to some of the early usage of CF [5]. Traditionally, CF is characteristic of its high dimensional real number vector space (i.e. $\Re^n$ for large integer n), and its symmetric correlation/measure for finding so-called neighbors [17]. These elements of CF have inspired us in creating our model, which inherits from CF and pushes it further. In our model, instead of the real space $\Re^n$, we venture to consider the problem in polynomial space P^n; instead of symmetric measure, we generalize to multiple asymmetric measurements on P^n. Indeed, there are studies on asymmetric similarity modeling [3] and on vector space modeling [7], but the setting is usually in $\Re^n$. Most of such studies come from clinical experiments in psychology and the modeling is often ad hoc [24].

3. MODEL
Figure 1 shows the overall design of our proposed recommendation system in a flow diagram. In this section, we first state our goal and then provide our solution with a mathematical model.

3.1 Our Goal
Given a SNS described above, we want, for a given product, an ordered list of recommendations based on high-level opinion and sentiment information (e.g. user preferences, attitude) extracted from the SNS.
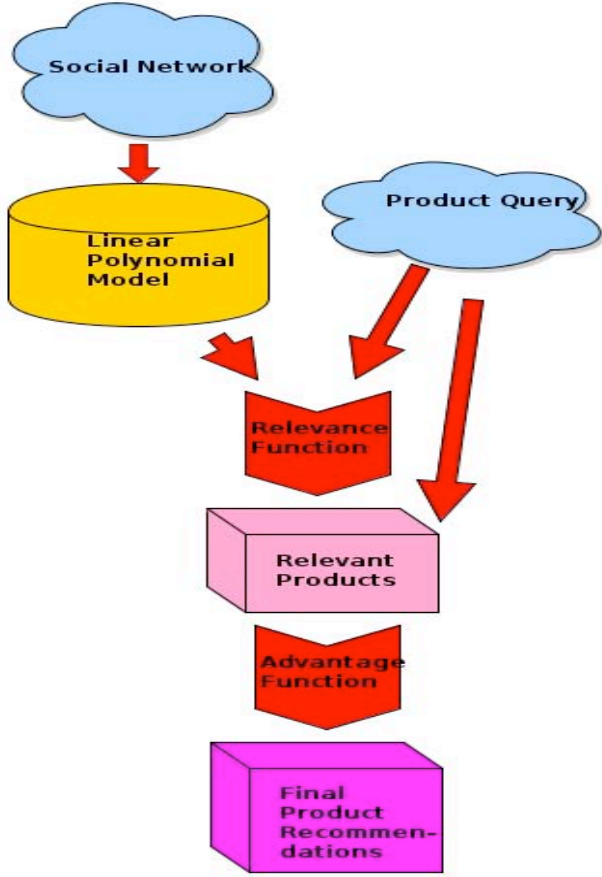
Figure 3. Overall flow diagram of the recommendation system.

3.2 Basic Model Definitions

We attempt to abstract the SNS into finite-dimensional linear spaces. We briefly describe each object in the SNS to be modeled and we motivate our definitions in the nontrivial cases. Then we give our definitions for similarity measurement and advantage measurement, the two measurements that are central to our system.

Let S denote the SNS, which is a countable collection of walls. $S = \{W_1, W_2, ..., W_{|S|}\}$. A wall on SNS, denoted by W, is a countable collection of comments. Further, the walls are pairwise disjoint and the union of all walls forms the collection of all comments: $W_i = \{c_{i,1}, c_{i,1}, ..., c_{i,|W_i|}\}$. The comments are unique across S, by which we mean that all the comments form a single collection C. Since any comment is created by a certain user and is posted to a certain wall, we preserve such associations as partitions on the set C. In particular, $\{W_1, W_2, ..., W_{|S|}\}$ is a countable partition on the set C based on the walls where $W_j \bigcap W_i = \varnothing$ for $\forall i \neq j$ and $\bigcup_{i=1}^{|S|} W_i = C$. Similarly, $\{u_1, u_2, ..., u_{|U|}\}$ is a countable partition on the set C based on the users. So far our definitions for S, W and C have been abstract. Defining the comment object allows us to substantiate the previous definitions. Most comments contain valuable information about the sentiment of the user towards the wall or the product/brand the wall represents. We represent the sentiment as an integer between -10 and 10 with -10 being the most negative sentiment and 10 being the most positive sentiment. For the derivation of this sentiment score, please refer to our previous work [18]. Besides its sentiment, the popularity of a comment also accounts for a vital part of the comment's role in the SNS. We measure the popularity of a particular comment by counting the number of "likes" it received from other users. So we represent a comment c simply as a tuple (s, v) where the sentiment $s \in \{-10, ..., 10\}$ and the popularity $v \in \{0, 1, 2, ...\}$. To model the SNS users, we let $U = \{u_1, ..., u_{|U|}\}$ be the set of all users on S. In an abstract way, we see the set of users as a partition on C. Furthermore, the user-partition on C is always finer than the wall-partition on C; so we may also use a subset of U to represent a wall W and denote the user-represented wall as $\tilde{W}$:

$$\tilde{W} = \inf\left\{u_i \in U : \bigcup_i u_i \supseteq W\right\}, \tag{1}$$

which can be intuitively understood as the minimal subset of U that covers all comments present on W.

A concrete representation of users plays a critical role in our system. Among the many possible forms in which we can represent a user, we think polynomial is an excellent mathematical description for a user in the situation of modeling a SNS for ranking purposes. Our justification for this choice is given in the following paragraph. Let

$$p_{u,W}(x) = \sum_v \left[\left(\sum_i^{|r_v|} s_{i,v}\right) r_v x^v\right] \tag{2}$$

be the representation of a user u's presence on a particular wall W where the integer index v is the popularity, $s_{i,v}$ is the sentiment of the i-th comment by user u on wall W with v likes from others, and r_v is the number of such comments by user u on wall W with v likes from others.

A polynomial (not the value of a polynomial) is fully determined by the coefficients and the exponents of its terms. We believe that interpreting popularity by the power law, sentiment by the sign of the polynomial's evaluation, and quantity of comments by scalar coefficients can adequately approximate and measure a user's influence and aggregated opinion in the SNS while taking advantage of the mathematical convenience of the polynomials. In fact, many preeminent previous works have benefited from similar assumptions [1].

3.3 Two Asymmetric Measurements

The Similarity Measurement: In a recommendation system, we often need to define some measurements to identify similarities, based on which an algorithm can cleverly identify so-thought similar items [2]. Many previous works extensively study symmetric similarity measurements such as Dice's coefficient, Jaccard's coefficient, cosine similarity, etc. [12]. However, these studies are hardly applicable to our situation because we think asymmetric similarity is characteristic of a social opinion-based system. For example, if an online customer is choosing an HDTV, then it makes sense to recommend HDTV connection cables for this customer, but on the other hand, if the customer is choosing from HDTV connection cables online, it would be less productive to recommend HDTVs to him/her: if someone is actively looking for an HDTV connector, the chances are that they already have an HDTV.

Our similarity measurement definition should:

1. Bear the intuition of product similarity,
2. Be asymmetric by default, and

3. Be mathematically convenient and reducible (because in the experiments, we would experience performance issues if not so).

Let P^m denote the vector space of all real polynomials with maximum degree m so that all the polynomials $p_{u,W}$ live in P^m.

Let $\langle p,q \rangle = \int_0^a pq$ for $\forall p,q \in P^m$ denote the conventional inner product on P^∞ and P^m. We propose a similarity measurement based on our requirements:

$$\langle\langle p,q \rangle\rangle = \frac{\langle p-q, p-q \rangle}{\exp(\langle q,q \rangle)} \text{ for } \forall p,q \in P^m, \tag{3}$$

where the numerator is the standard metric function for P^m and the denominator introduces asymmetry. We motivate the exponential asymmetry of the denominator in an example. Let us suppose that p is a user polynomial on W_A and q is a user polynomial on W_B. And suppose that q has some higher power terms with very positive coefficient, which makes $\exp(\langle q,q \rangle)$ large while p is a polynomial with lower power terms and small coefficients and has a small $\exp(\langle p,p \rangle)$. Now this example leads to a small $\langle\langle p,q \rangle\rangle$ and a large $\langle\langle q,p \rangle\rangle$, which means that p is more similar to q than q is similar to p. This result coincides with our daily experience with dominant object and dominated object: e.g. people tend to think that Vietnam is more similar to China than China is to Vietnam. Also notice that adjusting the parameter a in the definition for $\langle \cdot,\cdot \rangle$ allows us to weigh the exponents of the polynomial differently. Intuitively, the larger a is, the more weight we assign to higher power terms, i.e. comments with more likes. For positive a, $\langle\langle \cdot,\cdot \rangle\rangle$ is smaller if the two polynomials are more similar.

It turns out that our definitions for the similarity measurement, $\langle\langle \cdot,\cdot \rangle\rangle$, induces convenient properties, i.e. properties that can greatly reduce computational cost and make our large-scale experiment feasible to carry out. The first property we derive for the similarity measurement is the countable additivity in the first slot of $\langle\langle \cdot,\cdot \rangle\rangle$.

Let $a,b,c \in P^m$; then we have

$$\langle\langle a+b,c \rangle\rangle = \frac{\langle a+b-c, a+b-c \rangle}{\exp(\langle c,c \rangle)}$$

$$= \frac{\langle a-c, a+b-c \rangle + \langle b, a+b-c \rangle}{\exp(\langle c,c \rangle)}$$

$$= \frac{\langle a-c, a-c \rangle + \langle a-c, b \rangle + \langle b,a \rangle + \langle b, b-c \rangle}{\exp(\langle c,c \rangle)}$$

$$= \frac{\langle a-c,a-c \rangle + \langle a-c,b \rangle + \langle a,b \rangle + \langle c,b-c \rangle - \langle c,b-c \rangle + \langle b,b-c \rangle}{\exp(\langle c,c \rangle)}$$

$$= \langle\langle a,c \rangle\rangle + \frac{\langle a-c,b \rangle + \langle a,b \rangle + \langle c,b \rangle - \langle c,c \rangle + \langle b-c,b-c \rangle}{\exp(\langle c,c \rangle)}$$

$$= \langle\langle a,c \rangle\rangle + \frac{\langle a,b \rangle - \langle c,b \rangle + \langle a,b \rangle + \langle c,b \rangle - \langle c,c \rangle + \langle b-c,b-c \rangle}{\exp(\langle c,c \rangle)}$$

$$= \langle\langle a,c \rangle\rangle + \langle\langle b,c \rangle\rangle + \frac{2\langle a,b \rangle - \langle 0-c, 0-c \rangle}{\exp(\langle c,c \rangle)}$$

$$= \langle\langle a,c \rangle\rangle + \langle\langle b,c \rangle\rangle + \langle\langle 0,c \rangle\rangle + \frac{2\langle a,b \rangle}{\exp(\langle c,c \rangle)}. \tag{4}$$

Note that in the above derivation, we heavily rely on the properties of inner product $\langle \cdot,\cdot \rangle$ on P^m. Similarly for $a,b,c,d \in P^m$, we have:

$$\langle\langle a+b+c,d \rangle\rangle = \langle\langle a,d \rangle\rangle + \langle\langle b,d \rangle\rangle + \langle\langle c,d \rangle\rangle + \langle\langle 0,d \rangle\rangle + \frac{\langle a,b+c \rangle + \langle b,c \rangle}{\exp(\langle d,d \rangle)} \tag{5}$$

In fact, by the strong mathematical induction, for $p_1,...,p_N,q \in P^m$, we have:

$$\left\langle\left\langle \sum_{i=1}^N p_i, q \right\rangle\right\rangle = \sum_{i=1}^N \langle\langle p_i,q \rangle\rangle + \langle\langle 0,q \rangle\rangle + \frac{\sum_{i=1}^{N-1} \left\langle p_i, \sum_{j=i+1}^N p_j \right\rangle}{\exp(\langle q,q \rangle)}. \tag{6}$$

An important and immediate implication of the above formula is that if $p_1,...,p_N$ form a subset of an orthogonal basis for P^m, then the last term on the RHS vanishes due to orthogonality. Further, since the summation on the LHS is inside the $\langle\langle \cdot,\cdot \rangle\rangle$ operator, we can let $(o_1,...,o_m)$ be an orthogonal basis that spans P^m such that $\langle o_i, o_j \rangle = 0$ for $\forall i \neq j$ and $\sum_{i=1}^m o_i = \sum_{i=1}^N p_i$. It follows then:

$$\left\langle\left\langle \sum_{i=1}^N p_i, q \right\rangle\right\rangle = \left\langle\left\langle \sum_{i=1}^m o_i, q \right\rangle\right\rangle = \sum_{i=1}^m \langle\langle o_i,q \rangle\rangle + \langle\langle 0,q \rangle\rangle + \frac{\sum_{i=1}^{m-1} \left\langle o_i, \sum_{j=i+1}^m o_j \right\rangle}{\exp(\langle q,q \rangle)}$$

$$\left\langle\left\langle \sum_{i=1}^N p_i, q \right\rangle\right\rangle = \sum_{i=1}^m \langle\langle o_i,q \rangle\rangle + \langle\langle 0,q \rangle\rangle. \tag{7}$$

Further, for each i, we construct an orthogonal basis $(p_i, o_{i,1},...,o_{i,m-1})$ that spans P^m. This allows us to rewrite the first-slot additivity:

$$\left\langle\left\langle \sum_{i=1}^N p_i, q \right\rangle\right\rangle = \sum_{i=1}^N \langle\langle p_i,q \rangle\rangle + \langle\langle 0,q \rangle\rangle + \frac{\sum_{i=1}^{N-1} \left\langle p_i, c_{i,0}p_i + \sum_{j=1}^{m-1} c_{i,j}o_{i,j} \right\rangle}{\exp(\langle q,q \rangle)}$$

$$= \sum_{i=1}^N \langle\langle p_i,q \rangle\rangle + \langle\langle 0,q \rangle\rangle + \frac{\sum_{i=1}^{N-1} \langle p_i, c_{i,0}p_i \rangle}{\exp(\langle q,q \rangle)}$$

$$= \sum_{i=1}^N \langle\langle p_i,q \rangle\rangle + \langle\langle 0,q \rangle\rangle + \frac{\sum_{i=1}^{N-1} c_{i,0}\langle p_i,p_i \rangle}{\exp(\langle q,q \rangle)}, \tag{8}$$

where the $c_{i,j}$'s are the coefficients when we write $\sum_{j=i+1}^N p_j$ as a linear combination of the orthogonal basis $(p_i, o_{i,1},...,o_{i,m-1})$.

Now we show that construction of such orthogonal bases is always possible as long as P^m is finite dimensional. Let B be the list $(x^0, x^1, x^2,...,x^m)$. Clearly B is a basis for P^m. Now for any p_i, let n be the smallest number such that p_i's n-th power term has a non-zero coefficient. Then we construct a list $B' = (x^0,...,x^{n-1}, p_i, x^{n+1},...,x^m)$. Clearly B' is also a basis for P^m. Then it follows from the Gram-Schmidt process that an orthogonal basis $(p_i, o_{i,1},...,o_{i,m-1})$ can always be constructed from B'.

A special case for first slot additivity leads us to the second property: integer multiplication in the first slot. Note that for the purpose of our model, integer multiplication is sufficient since all the coefficients and exponents in the polynomials are integers by our definition. For $p, q \in P^m$ and $n = 2, 3, \ldots$, we have:

$$\langle\langle np, q \rangle\rangle = n \langle\langle p, q \rangle\rangle + \langle\langle 0, q \rangle\rangle + \frac{\sum_{i=1}^{n-1} \langle p, (n-i) p \rangle}{\exp(\langle q, q \rangle)}$$

$$= n \langle\langle p, q \rangle\rangle + \langle\langle 0, q \rangle\rangle + \frac{\sum_{i=1}^{n-1} (n-i) \langle p, p \rangle}{\exp(\langle q, q \rangle)}$$

$$= n \langle\langle p, q \rangle\rangle + \langle\langle 0, q \rangle\rangle + \frac{\langle p, p \rangle \left(\sum_{i=1}^{n-1} n - \sum_{i=1}^{n-1} i \right)}{\exp(\langle q, q \rangle)}$$

$$= n \langle\langle p, q \rangle\rangle + \langle\langle 0, q \rangle\rangle + \frac{\langle p, p \rangle \left(\sum_{i=1}^{n-1} n - \sum_{i=1}^{n-1} i \right)}{\exp(\langle q, q \rangle)}$$

$$= n \langle\langle p, q \rangle\rangle + \langle\langle 0, q \rangle\rangle + \frac{\langle p+q-q, p+q-q \rangle \left(\frac{n(n-1)}{2} \right)}{\exp(\langle q, q \rangle)}$$

$$= n \langle\langle p, q \rangle\rangle + \langle\langle 0, q \rangle\rangle + \left(\frac{n^2 - n}{2} \right) \langle\langle p+q, q \rangle\rangle$$

$$= n \langle\langle p, q \rangle\rangle + \langle\langle 0, q \rangle\rangle + \left(\frac{n^2 - n}{2} \right) \left(\langle\langle p, q \rangle\rangle + \langle\langle q, q \rangle\rangle + \frac{2\langle p, q \rangle}{\exp(\langle q, q \rangle)} - \langle\langle 0, q \rangle\rangle \right), \quad (9)$$

where we note that $\langle\langle q, q \rangle\rangle = 0$ for $\forall q \in P^m$

$$= \left(\frac{n^2 + n}{2} \right) \langle\langle p, q \rangle\rangle - \left(\frac{n^2 - n - 2}{2} \right) \langle\langle 0, q \rangle\rangle + \left(n^2 - n \right) \frac{\langle p, q \rangle}{\exp(\langle q, q \rangle)}. \quad (10)$$

Integer multiplication in the first slot enables us an alternative reduction for (0):

$$\left\langle\left\langle \sum_{i=1}^{N} p_i, q \right\rangle\right\rangle = \left\langle\left\langle \sum_{i=1}^{m} a_i x^i, q \right\rangle\right\rangle = \sum_{i=1}^{m} \left\langle\left\langle a_i x^i, q \right\rangle\right\rangle + \langle\langle 0, q \rangle\rangle + \frac{\sum_{i=1}^{m-1} \left\langle a_i x^i, \sum_{j=1}^{m-1} a_j x^j \right\rangle}{\exp(\langle q, q \rangle)}, \quad (11)$$

where a_i's are all integers by our assumption. And this form can be further reduced by our formula for integer multiplication in the first slot. This reduction by integer multiplication has the same time complexity as the reduction by constructing orthogonal bases in terms of the number of times when numeric integrals are evaluated.

We can also derive the second slot additivity; however, the additivity for the second slot for $\langle\langle \cdot, \cdot \rangle\rangle$ is hardly reducible and does not benefit us in an immediate way. Thus, we do not include that part here.

The Advantage Measurement: Unlike a classification task, a recommendation system not only needs a similarity measurement, but it also requires an advantage measurement, a measurement by which it can rank the identified similar items. Ideally according to the advantage measurement, an item should get a positive/negative measurement if it has an advantage/disadvantage over another item, based on related social opinions. In order to give productive recommendations, we first need to understand the customers, to understand why customers are looking for certain products online. In fact, field experiments show that a large portion of online customers shop online because they are unhappy

with a current product and are looking for an alternative [6]. For example, if a customer is unhappy with her Dell laptop and is looking for a replacement from Apple, then it makes little sense to recommend Dell products, even though, Dell is identified as similar enough to Apple, which would make Dell recommendable if we only considered relevance in our system. Brand loyalty is another aspect where advantage measurement comes into play. For example, it is almost counter-productive to recommend Nikon products to a loyal Canon camera customer, although Nikon and Canon are probably very close to each other in the sense of similarity.

However, it is not easy to identify things like an unhappy Dell user or the average loyalty level among Canon users, etc. from a swamp of social opinions. We think that understanding the high-level meanings of user-generated conversations on the SNS is key to establishing an advantage measurement that can handle described situations. We define the advantage measurement function $\langle | \cdot, \cdot | \rangle$ as $\langle | p, q | \rangle = \int_0^a p^+ \cdot q^- - q^+ \cdot p^-$ for $\forall p, q \in P^m$ where $p^+ = \max(0, p)$ and $p^- = -\min(0, p)$. $\langle | p, q | \rangle$ can be interpreted as "p's advantage over q". For example, let us suppose that p and q are the polynomials of a same user on two different walls and the user leaves more positive comments on the p-wall than she does for the q-wall. Then as a result, $p^+ \cdot q^-$ becomes greater than $q^+ \cdot p^-$ and therefore $\langle | p, q | \rangle$ evaluates to be positive, which is understood as p-wall's advantage over q-wall in this user's opinion. In fact, in our application the two arguments of $\langle | \cdot, \cdot | \rangle$ are always going to be the two user polynomials for the same user on two different walls, although its definition we give is more general.

Here we try to investigate arithmetic properties of $\langle | \cdot, \cdot | \rangle$. First, the commutative property is obvious: $\langle | p, q | \rangle = -\langle | q, p | \rangle$ for $\forall p, q \in P^m$

However, there is little that we can induce for additivity in the first slot due to the inconvenient fact that $(p+q)^+ = p^+ + q^+ + (p^+ - q^-)^+ + (q^+ - p^-)^+$. Fortunately, in the application and experiments we show that we do not need nice additivity properties for reducing the computational cost, as much we do need for the $\langle\langle \cdot, \cdot \rangle\rangle$ operator.

3.4 Modeling Product Relevance

Suppose we have two products A and B; and each corresponds to only one wall (W_A and W_B respectively) on the social network. Then we calculate their relevance $R(\tilde{W}_A, \tilde{W}_B)$ by:

$$R(\tilde{W}_A, \tilde{W}_B) = \sum_{i=1}^{|\tilde{W}_A|} \sum_{j=1}^{|\tilde{W}_B|} \langle\langle p_{i,W_A}, q_{j,W_B} \rangle\rangle, \text{ where } p_i \in \tilde{W}_A \text{ and } q_i \in \tilde{W}_B. \quad (12)$$

However, computing the relevance in the above definition is very costly due to the huge SNS and the integral-based definition for $\langle\langle \cdot, \cdot \rangle\rangle$. Now we show how to simplify the computation.

For a fixed j, consider the sum $\sum_{i=1}^{|\tilde{W}_A|} \langle\langle p_{i,W_A}, q_{j,W_B} \rangle\rangle$. It follows from section 3.3 that

$$\sum_{i=1}^{|\tilde{W}_A|} \left\langle\!\left\langle p_{i,W_A}, q_{j,W_B} \right\rangle\!\right\rangle = \left\langle\!\left\langle \sum_{i=1}^{|\tilde{W}_A|} p_i, q_j \right\rangle\!\right\rangle - \left\langle\!\left\langle 0, q_j \right\rangle\!\right\rangle - \frac{\sum_{i=1}^{|\tilde{W}_A|} c_{i,0}\left\langle p_i, p_i \right\rangle}{\exp\left(\left\langle q_j, q_j \right\rangle\right)}. \tag{13}$$

Then we can rewrite:

$$\mathrm{R}\left(\tilde{W}_A, \tilde{W}_B\right) = \sum_{j=1}^{|\tilde{W}_B|} \left(\left\langle\!\left\langle \sum_{i=1}^{|\tilde{W}_A|} p_i, q_j \right\rangle\!\right\rangle - \left\langle\!\left\langle 0, q_j \right\rangle\!\right\rangle - \frac{\sum_{i=1}^{|\tilde{W}_A|} c_{i,0}\left\langle p_i, p_i \right\rangle}{\exp\left(\left\langle q_j, q_j \right\rangle\right)} \right)$$

$$= \sum_{j=1}^{|\tilde{W}_B|} \left\langle\!\left\langle \sum_{i=1}^{|\tilde{W}_A|} p_i, q_j \right\rangle\!\right\rangle - \sum_{j=1}^{|\tilde{W}_B|} \left\langle\!\left\langle 0, q_j \right\rangle\!\right\rangle - \sum_{j=1}^{|\tilde{W}_B|} \left(\frac{\sum_{i=1}^{|\tilde{W}_A|} c_{i,0}\left\langle p_i, p_i \right\rangle}{\exp\left(\left\langle q_j, q_j \right\rangle\right)} \right)$$

$$= \sum_{j=1}^{|\tilde{W}_B|} \left\langle\!\left\langle \sum_{i=1}^{|\tilde{W}_A|} p_i, q_j \right\rangle\!\right\rangle - \sum_{j=1}^{|\tilde{W}_B|} \left\langle\!\left\langle 0, q_j \right\rangle\!\right\rangle - \left(\sum_{i=1}^{|\tilde{W}_A|} c_{i,0}\left\langle p_i, p_i \right\rangle \right)\left(\sum_{j=1}^{|\tilde{W}_B|} \frac{1}{\exp\left(\left\langle q_j, q_j \right\rangle\right)} \right). \tag{14}$$

The above expression reduces the computational cost of $\mathrm{R}\left(\tilde{W}_A, \tilde{W}_B\right)$ from $O\left(|\tilde{W}_A|\|\tilde{W}_B|\right)$ to $O\left(|\tilde{W}_A| + |\tilde{W}_B|\right)$ in terms of the times when numeric integration is performed. In order to apply this linear running time formula for calculating $\mathrm{R}\left(\tilde{W}_A, \tilde{W}_B\right)$, we do need additional pre-calculation for each $c_{i,0}$ and $\left\langle p_i, p_i \right\rangle$, for which we first need to construct an orthogonal basis for each p_i. Fortunately all such pre-calculation is linear in running time.

3.5 Modeling Product Advantage

Continuing our running example from previous section 3.4, we measure the advantage of product A over product B by the following:

$$A\left(\tilde{W}_A, \tilde{W}_B\right) = \sum_{i=1}^{|\tilde{W}_{AB}|} \left\langle \left| p_{i,W_A}, p_{i,W_B} \right| \right\rangle, \text{ where } \tilde{W}_{AB} = \tilde{W}_A \cap \tilde{W}_B \text{ and } p_i \in \tilde{W}_{AB}. \tag{15}$$

Intuitively, this advantage function formula means the summation of their common users' opinions on the two comparing walls. Due to $\left|\tilde{W}_{AB}\right| << \left|\tilde{W}_A \cup \tilde{W}_B\right|$ in practice, the calculation of this advantage function is much faster than that of the relevance function. Finally, we take a weighted average of $\mathrm{R}\left(\tilde{W}_A, \tilde{W}_B\right)$ and $A\left(\tilde{W}_A, \tilde{W}_B\right)$ to achieve our final score, which we call "AdvR" in our experiments. Note: we have to reversely re-assign scores for $\mathrm{R}\left(\tilde{W}_A, \tilde{W}_B\right)$ because in the original definition, smaller $\left\langle\!\left\langle .,. \right\rangle\!\right\rangle$ means higher relevance.

4. EXPERIMENTS

In this section, we first present a detailed description of our experiments, which includes the dataset used and evaluation methods. Then, we analyze the performance results. Finally, we present two case studies that validates our intuition and real-life experience regarding what products should be recommended for a given product.

4.1 The Datasets

The FB-8000 is a large replication of Facebook's public wall data. Specifically, FB-8000 contains 8094 Facebook public walls, of which almost all are the most popular and active public walls on Facebook and all are synchronized with Facebook within real-time latency. FB-8000 also contains over 196 million unique users, over 360 million unique comments, and over 1.83 billion unique "likes".
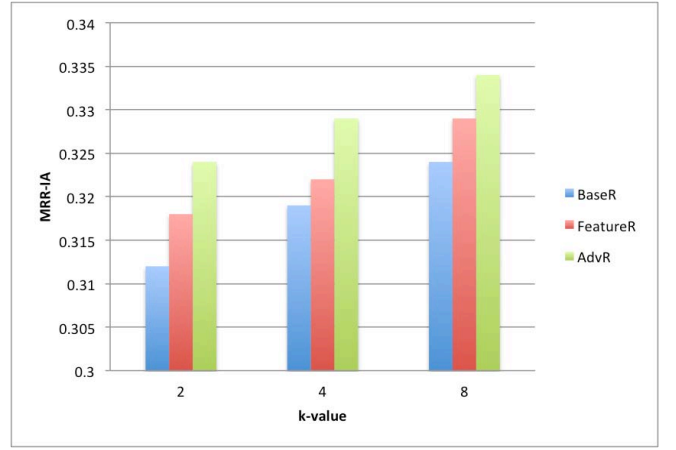

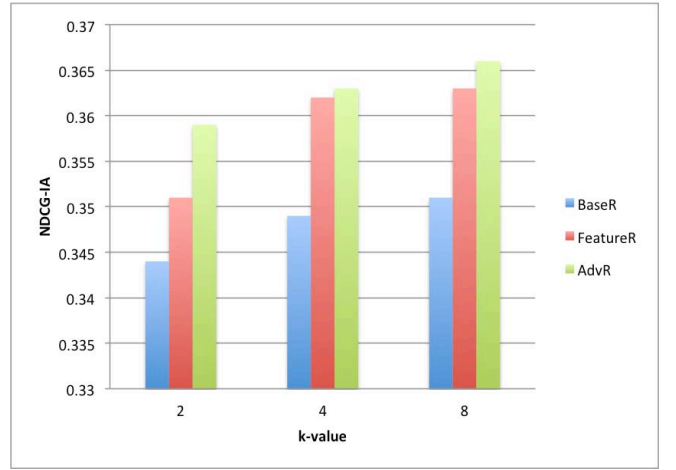**Figure 4. MRR-IA scores for three methods.**

Figure 5. NDCG-IA scores for three methods.

The AMZN-900 is a replication for the products for sale via Amazon.com and their associated RRN. AMZN-900 contains 914 different products, over 330 thousand unique reviewers, and over 510 thousand unique reviews and ratings. This dataset represents the traditional structure, from which product recommendations are made. We use AMZN-900 for comparison in the first case study.

The LNKD-2000 is a replication of the company public profiles on LinkedIn network. LNKD contains 571 company profiles, over 1.5 million unique user profiles, over 3.1 million unique comments and over 1.1 million unique "likes". Clearly, LNKD-2000 does not fit into product ranking. However, we include this dataset to show how well our ranking approach can be generalized to other domains.

For each of the three datasets, we perform basic spam user filtering by excluding the users and their content if a user makes a large amount of comments with identical contents.

4.2 Results and Analysis

Since our recommendation system considers more feature dimensions than just relevance, we regard these features as diversification on top of the relevance ranking. Therefore, we use NDCG-IA (intent-aware normalized discounted cumulative gain) and MRR-IA (intent-aware mean reciprocal rank) as the evaluation metrics, which are proposed by [20], [21], and [22], and illustrated by [19].
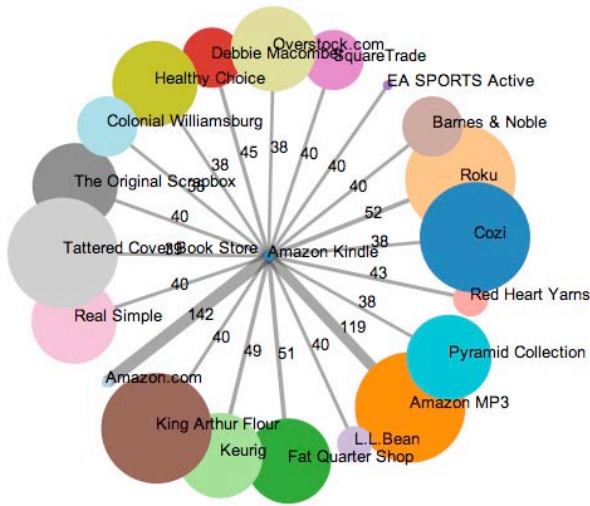
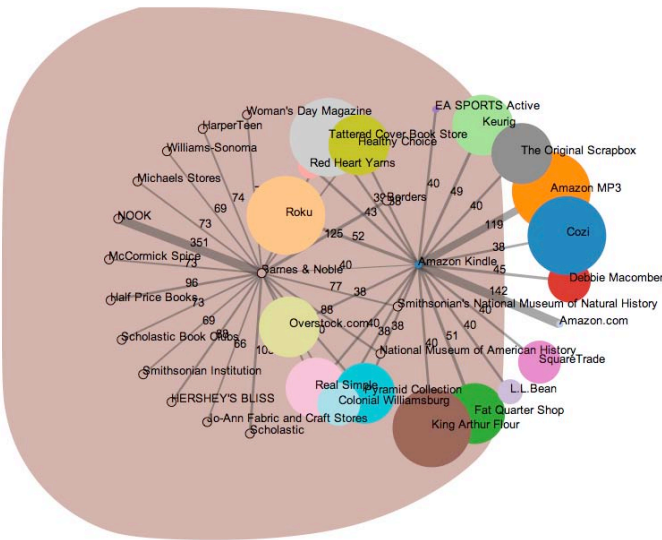**Figure 6. Top 20 recommendations for Amazon Kindle.**

Figure 7. Breadth-first expansion on Barnes & Noble for more recommendations for Amazon Kindle.

Outside the proposed system, we include two more methods in our experiments: baseline relevance rank, and feature rank from our previous work [15]. Our baseline relevance is simply the rank of products in the descending order of relevance without considering any user sentiment or user opinion, where relevance is quantified as: $\mathrm{BaseR}\left(\tilde{W}_A, \tilde{W}_B\right) = |\tilde{W}_A \cap \tilde{W}_B|$, where A is the fixed product, the product whose relevant items, represented as B's, are to be recommended. Conceptually, BaseR is simply the descending order of the number of common Facebook users. In Figure 4 and Figure 5, we plot the performance of the three methods for three different k-values. The baseline relevance is labeled as "BaseR"; the proposed method is labeled as "AdvR"; and the feature rank is labeled as "FeatureR". It is clear from both Figure 4 and Figure 5 that MRR-IA scores and NDCG-IA scores increase as k-value increases for all three methods. Feature ranking gives a significant gain in both scores over the baseline relevance and the proposed method is comparable with the feature rank. Note that the results for feature rank (and therefore the comparing results in Figure 4 and Figure 5) are produced based on AMZN-900 dataset, which is a RRN with structured information.

This means that using external SSN like Facebook, our proposed recommendations can perform comparably with a state-of-the-art ranking system that uses well-structured RRN information.

4.3 Case Studies

Our first case study shows how the recommendations really look when we apply our method to a particular product in the FB-8000 dataset. We choose the Amazon Kindle as our focal product. We visualize its top-20 recommendations in Figure 6. In Figure 6, Amazon Kindle is centralized as a dot with 20 connecting edges to 20 recommended products. The number on each edge and the thickness of each edge represent a normalized combination of relevance score and advantage score. The color scheme for each recommended product is assigned randomly but the size of its circle is determined by the statistical confidence in recommending it. On the recommendation list, we note that many items on the list are highly relevant to the focal product and most of them, such as Amazon MP3, receive generally positive opinions from the public (and a sizeable volume of such comments which results in its high statistical confidence) while the not-so-successful similar products, such as the HP TouchPad, are not found on the list.

For expanding the list beyond the top 20, we develop so-called "breadth-first" expansion instead of going further down Amazon Kindle's list for the next 20-40 items. For example, when we run our recommendation system on "Barnes & Noble", one of Kindle's recommendation, we get interesting and relevant results as shown in Figure 7. This concept of "breadth-first" expansion is useful and remediate in practice. For example, the product NOOK, highly relevant and competitive to Kindle, is successfully reached by the breadth-first expansion on "Barnes & Noble"; NOOK would not otherwise show on the list even in the top 40 recommendations for Kindle (we believe that the main reason NOOK fails to earn a high spot on Kindle's list is because of Kindle fans' "loyalty effect" acting strongly on the advantage score, which is the dominant score in our ranking scheme after relevant products are identified).

Table 2. Top Career Recommendations from LinkedIn.

Top 5 Career Recommendations for IBM			
Company	**Field**	**Relevance**	**Advantage**
BT	IT	50	82
Nielson	Measurement	48	74
Standard Chartered	Finance	40	75
Booz Allen Hamilton	Consulting	33	66
State Farm	Insurance	31	58
Top 5 Career Recommendations for ESPN			
Hertz	Car Rental	28	55
CBRE	Real Estate	32	41
Oracle	IT	29	30
Northern Trust	Finance	18	31
Marsh	Insurance	16	31
Top 5 Career Recommendations for Groupon			
Novartis	Pharmaceutical	45	66
Gap	Retail	44	59
MSNBC	Media	32	50
Capgemini	Consulting	27	52
GSK	Pharmaceutical	29	41

Second, we try to apply our method to a non-product recommendation task: job recommendation. We conduct this experiment on the LNKD-2000 dataset with IBM, ESPN, and Groupon as our choice of focal companies. Table 2 summarizes our findings. We choose these three companies because IBM is very different from Groupon in size and ESPN is very different from the other two in field. Despite these disparities in the focal companies, we observe interesting similarities in their recommendations. For example, both IBM and ESPN have an IT company, an insurance company, and a finance company in their top 5 lists. Our case study also shows great diversity in each of the lists: the 15 recommended companies we list represent 10 different fields; and each top 5 list contains at least 4 fields. We believe our modeling approaching is, in a good way, highly invariant to the size of the dataset because in our experiments FB-8000 and LNKD-2000 have big differences in terms of data size.

5. CONCLUSION AND FUTURE WORK

Intelligent recommendation systems have become more and more important as e-commerce customers face a deluge of options. This topic also attracts a lot of interest from academia. In this paper, we present a novel approach for building a recommendation system based on external SNS opinion and sentiment information, which is modeled in a polynomial vector space. In the modeling process, we define and rationalize two concepts of measurements: the similarity measurement and the advantage measurement. Later in application, we demonstrate how to reduce the computational cost of using this model by exploiting special properties of the measurements that we define. Experiments on large, real-world datasets, including two case studies, show that our system yields promising results, which establishes our model as a promising application of sentiment analysis results.

In the future, we want to provide personalized results, although such ideas require non-trivial privacy-preserving experiment design. Also, focusing on innovative methods of incorporating high-level meanings more than text sentiment can further advance our system.

6. ACKNOWLEDGEMENT

7. REFERENCES

[1] R. Malmgren, D. Stouffer, A. Campanharo, and L. Amaral *Science* 325, 1696-1700. 2009

[2] P. Symeonidis, E. Tiakas, and Y. Manolopoulos. Product recommendation and rating prediction based on multi-modal social networks. In *RecSys '11*. ACM, 2011

[3] J. He and W. C. Wesley. A Social Network-Based Recommender System (SNRS). *Data Mining for Social Network Data* 12: 47–74. 2010

[4] M. Jamali and M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *RecSys '10*. ACM, 2010

[5] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Analysis of recommendation algorithms for e-commerce. In *EC '00*. ACM, 2000

[6] Zettelmeyer, Florian, F. S. Morton, and J. Silva-Risso. *Journal of Marketing Research* 43, no. 2. 2006.

[7] E. Cambria, T. Mazzocco, A. Hussain, C. Eckl: Sentic Medoids: Organizing Affective Common Sense Knowledge in a Multi-Dimensional Vector Space. *ISNN '11*. 2011

[8] Q. Yuan, L. Chen, and S. Zhao. Factorization vs. regularization: fusing heterogeneous social relationships in top-n recommendation. In *RecSys '11*. ACM, 2011

[9] S. Isaacman, S. Ioannidis, A. Chaintreau, and M. Martonosi. Distributed rating prediction in user generated content streams. In *RecSys '11*. ACM, 2011

[10] S. Castagnos and A. Boyer. Personalized communities in a distributed recommender system. In *ECIR'07*. 2007

[11] K. Liu, K. Bhaduri, K. Das, P. Nguyen, and H. Kargupta. Client-side web mining for community formation in peer-to-peer environments. In *WEBKDD '06*. ACM, 2006

[12] J. Chen, J. Fagnan, R. Goebel, R. Rabbany, F. Sangi, M. Takaffoli, E. Verbeek, O. Zaïane, Meerkat: Community Mining with Dynamic Social Networks, *ICDMW '11*. 2011

[13] Maclean, Diana, and M. S. Lam. All Friends are Not Equal: Using Weights in Social Graphs to Improve Search. *Social Networks*. 2010

[14] K. Zhang, R. Narayanan, and A. Choudhary. Voice of the customers: mining online customer reviews for product feature-based ranking. In *WOSN'10*. USENIX, 2010

[15] K. Zhang, Y. Cheng, W.-k. Liao, and A. Choudhary. Mining Millions of Reviews: A Technique to Rank Products Based on Importance of Reviews, *ICEC*. 2011

[16] S. Huang, D. Shen, W. Feng, C. Baudin and Y. Zhang. Discovering Clues for Review Quality from Author's Behaviors on E-commerce Sites. *Pacific Asia Journal of the Association for Information Systems*. 2009

[17] T. Zhang and V. S. Iyengar. Recommender systems using linear classifiers. *J. Mach. Learn. Res.* 2 (March 2002). 2002.

[18] K. Zhang, Y. Cheng, Y. Xie, A. Agrawal, D. Palsetia, K. Lee, and A. Choudhary. SES: Sentiment Elicitation System for Social Media Data, *ICDM-SENTIRE '11*. IEEE, 2011

[19] X. Chen, H. Wang, X. Sun, J. Pan, and Y. Yu. Diversifying product search results. In *SIGIR '09*. ACM, 2011

[20] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM '09*. ACM, 2009

[21] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98*. ACM, 1998

[22] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *WWW '09*. ACM, 2009

[23] M. Goldberg, S. Kelley, M. Magdon-Ismail, K. Mertsalov, A. Wallace, Finding Overlapping Communities in Social Networks, *Social Computing (SocialCom) '10. IEEE*, 2010

[24] M. Johannesson. Modelling asymmetric similarity with prominence. *The British journal of mathematical and statistical psychology* 53 (Pt 1), no. 1991: 121-139. 2000