Social Group Modeling with Probabilistic Soft Logic

Bert Huang, Stephen H. Bach, Eric Norris, Jay Pujara, Lise Getoor
Department of Computer Science
University of Maryland
College Park, MD 20742
{bert, bach, enorris, jay, getoor}@cs.umd.edu

Abstract

In this work, we show how to model the group affiliations of social media users using probabilistic soft logic. We consider groups of a broad variety, motivated by ideas from the social sciences on groups and their roles in social identity. By modeling group affiliations, we allow the possibility of efficient higher-level relational reasoning about the groups themselves, where the number of groups is relatively small compared to the number of users. We discuss preliminary results from experiments using real social media data collected from Twitter.

1 Introduction

Many recent advances develop methods for analyzing and understanding huge amounts of social media data. Much of this effort aims to find novel techniques and systems to increase the efficiency of massive-scale, fine-grained reasoning at the level of user-nodes (Lin and Kolcz, 2012; Low et al., 2010, 2012). In this work, we explore a different yet complementary approach for scalable social media analysis. We show how to construct probabilistic models that aggregate users into groups, which allows more complex relational reasoning at the group-level by limiting the finer-grained reasoning to only consider group membership. In addition to the computational benefits of group modeling, the idea that individuals strongly identify by their association with groups has a long history in social psychology and sociology.

Modeling group affiliations probabilistically provides a principled way to handle the ambiguity of the group concept. Groups and their role in society are difficult to define, but we can reason about them probabilistically by identifying user behaviors we expect to see in social media data. We can then infer group affiliations as latent information from the observed data.

We present preliminary experiments using social media data from Twitter, modeling groups using *probabilistic soft logic* (Broecheler et al., 2010), a declarative language for relational probabilistic modeling. One advantage of using a declarative modeling language is the extensibility of the resulting models. One can seamlessly integrate a more complex relational model about the groups themselves into the group model we present here, thus producing a joint model at multiple levels of granularity.

1.1 Groups in the Social Sciences

Social identity theory (Hogg, 2006) attributes a significant portion of individuals' identities to their group affiliations. In particular, the theory distinguishes *social identity*, which is shared among members of the same group, from *personal identity*, which is unique to each individual. Intergroup relations are also well studied in the literature of social identity theory (Tajfel et al., 1971). In future work, we plan to extend the group affiliation model we develop in our preliminary experiments to include intergroup relational reasoning. We expect that the joint modeling of group affiliations with intergroup relations will yield a rich model that remains tractable in practice for large data sources.

Among many ideas relating social sentiment to group affiliation, the psychological phenomenon of *groupthink*, in which groups will form opinions or decide on actions by seeking uniformity instead of making more rational individual decisions, has been studied in various social sciences (Turner and Pratkanis, 1998). Additionally, formal definitions of sociological groups exist. For example, in addition to the involvement of groups in social identity theory, a classical definition by Cooley (1983) distinguishes *primary groups*, which are formed by face-to-face, personal relationships, from *secondary groups*, which are formed according to shared interests or traits, and may include members who have never actually met each other.

While these sociological and psychological ideas are constantly evolving in their respective scientific fields, the basic common-sense ideas about groups sharing common sentiment, individuals' behavior correlating with their group affiliations, and a higher-level intergroup social interaction motivate our modeling approach.

1.2 Related Work

In our empirical evaluation, we analyze Twitter data, with a focus on using the *hashtags* added by users to their tweets. Social tagging and Twitter hashtags in particular have been studied extensively in social media research. For instance, Yang et al. (2012) recently analyzed the dual-role of hashtags, where hashtags serve as both content bookmarks, as well as symbols of community membership. Using this analysis, they developed methods to predict, or recommend, hashtags based on tweet content. Chang (2010) analyzed hashtag usage to provide insight in how information diffuses through the social network. Multiple studies (Backstrom et al., 2006; Zheleva et al., 2009) model group evolution in social networks and provide analysis of observed group statistics in various data sets. In particular, these analyses provide insight into the behavior of groups over time.

To probabilistically model group membership, we use *probabilistic soft logic* (PSL), a declarative language for relational probability distributions. PSL uses *first-order logic* (FOL) syntax to define constraints and potential functions in a graphical model over the truth values of logical atoms. These truth values are relaxed to soft-logic, which allows joint *most probable explanation* (MPE) inference in PSL to be a convex optimization. The next section provides more details about PSL, as well as how we use PSL to model group affiliation.

2 Modeling Groups with Probabilistic Soft Logic

This section reviews details on probabilistic soft logic (PSL), how it converts its declarative syntax into a probabilistic model, how it performs inference in that model, and the specific model we use to infer group affiliation.

2.1 Probabilistic Soft Logic

Probabilistic soft logic (PSL) is a system for probabilistic modeling using first-order logic syntax. PSL uses soft truth values, relaxing truth to the interval $[0, 1]$ and adapts logical connectives accordingly. As a consequence of the soft logic formulation and the design of the PSL language, inference in PSL is a convex optimization problem. Additionally, the soft truth values allow the natural integration of external functions ranging in the same interval, such as normalized similarity functions. This section provides a short overview of PSL, its usage, and its internal representation.

PSL uses a syntax based on first-order logic (FOL) as its underlying modeling language. In a PSL program, relationships and attributes are modeled by user-defined *predicates* (of arbitrary arity), and first order *rules* model dependencies or constraints on these predicates. Each PSL rule's antecedent is a conjunction of atoms and its consequent is a disjunction. PSL rules can be assigned a weight from $\mathbb{R}^+$. A PSL program thus consists of a set of predicates, weighted rules involving these predicates, and known truth values of ground atoms derived from observed data. Inference for the PSL program is over the remaining, unknown truth values.

To mathematically represent soft logic, PSL uses the *Lukasiewicz t-norm* and its corresponding *co-norm* as the relaxation of the logical AND and OR, respectively. These relaxations are exact at the extremes, when variables are either true (1.0) or false (0.0), and provide a consistent mapping for values in-between. The formulas for the relaxation of the logical conjunction ($\wedge$), disjunction ($\vee$), and negation ($\neg$) are as follows:

$$a \,\tilde{\wedge}\, b = \max\{0, a + b - 1\},$$
$$a \,\tilde{\vee}\, b = \min\{a + b, 1\},$$
$$\tilde{\neg}\, a = 1 - a,$$

where we use ˜ to indicate the relaxation from the Boolean domain. Rules are evaluated using the Lukasiewicz norms by converting the implication operator with the identity

$$X \,\tilde{\Rightarrow}\, Y \equiv \tilde{\neg}\, X \,\tilde{\vee}\, Y.$$

The probability distribution defined by a PSL program measures the overall *distance to satisfaction*, which is a function of all ground rules' truth values. The more groundings of rules have high truth values in an interpretation, the more likely that interpretation is. The resulting distribution is the weighted log-linear distribution over ground rule potentials.

Considering each grounded rule a factor and each truth value a variable, this probability distribution is a Markov random field over continuous variables. Maximum likelihood inference for the unknown truth values corresponds to solving a linear program, where the truth-value variables are constrained to be consistent with respect to the t-norms and are weighted by rule potentials. Recently, Bach et al. (2012) developed a fast inference algorithm using *consensus optimization* allows inference with an order of magnitude speedup in practice. Additional details, including a description of a learning algorithm for setting the weights, are provided by Broecheler et al. (2010).

2.2 PSL Model for Group Membership

We consider the setting common in social media analysis where we have data consisting of a set of users, their posts, and messages to other users. Each post or message can be tagged with free-form, self-organized textual tags. For instance, in our experiments, we consider Twitter users and their tweets. Tweets can include other Twitter usernames, in which case that tweet is called a "mention". The tweets are tagged with "hashtags", which are tokens beginning with the # symbol (e.g., `#NIPS2012`). Additionally, each post can have an attached sentiment score, which is, for example, computed automatically or produced by user voting.

Predicates for our proposed group model are as follows: $\textsc{Posted}(U, P)$ indicates that user U posted P, $\textsc{MessageTo}(P, U)$ indicates that post P is a message to user U, $\textsc{Tagged}(P, T)$ indicates that post P is tagged with tag T. The sentiment of post P is modeled by predicate $\textsc{Positive}(P)$ and $\textsc{Negative}(P)$. For example, let P be post "`@berty38: Really looking forward to seeing Lake Tahoe for the first time #NIPS2012 @NipsConference`". Then the information in P could be encoded with:

1.0	:	$\textsc{Posted}$(`@berty38`, P)
1.0	:	$\textsc{MessageTo}$(P, `@NipsConference`)
0.9	:	$\textsc{Positive}(P)$
0.0	:	$\textsc{Negative}(P)$
1.0	:	$\textsc{Tagged}$(P, `#NIPS2012`).

Note that in the above example, we include the false (0.0) NEGATIVE predicate for completeness, though PSL uses a closed-world assumption, so in practice one does not need to enumerate false statements.

The previously defined predicates will be fully observed in our experimental setup. We also reason about (mostly) unobserved, latent predicates, which will be inferred. The latent group affiliations are represented by the predicate $\textsc{MemberOf}(U, G)$, which indicates that user U is a member of group G. We additionally model group sentiment toward topics by inferring predicates $\textsc{Likes}(G, T)$ and $\textsc{Dislikes}(G, T)$, which encode group G's attitude toward tag T.

From these predicates, we write rules that encode the ideas that: (1) users that message one another are likely to share group memberships, and (2) members of a group share common sentiment toward topics. The following rules encode the propagation of group affiliations through messages:

$$\textsc{MemberOf}(A, G) \,\tilde{\wedge}\, \textsc{Posted}(A, P) \,\tilde{\wedge}\, \textsc{MessageTo}(P, B) \,\tilde{\wedge}\, \textsc{Positive}(P) \,\tilde{\Rightarrow}\, \textsc{MemberOf}(B, G)$$
$$\textsc{MemberOf}(A, G) \,\tilde{\wedge}\, \textsc{Posted}(B, P) \,\tilde{\wedge}\, \textsc{MessageTo}(P, A) \,\tilde{\wedge}\, \textsc{Positive}(P) \,\tilde{\Rightarrow}\, \textsc{MemberOf}(B, G).$$

We include the POSITIVE predicate to filter out negative messages from this rule, since users who message each other with negative sentiment may be attacking one another, and thus are unlikely to share group affiliations.

The following rules encode the shared sentiment within groups:

$$\text{POSTED}(U,P) \tilde{\wedge} \text{TAGGED}(P,T) \tilde{\wedge} \text{POSITIVE}(P) \tilde{\wedge} \text{LIKES}(G,T) \tilde{\Rightarrow} \text{MEMBEROF}(U,G)$$
$$\text{POSTED}(U,P) \tilde{\wedge} \text{TAGGED}(P,T) \tilde{\wedge} \text{NEGATIVE}(P) \tilde{\wedge} \text{DISLIKES}(G,T) \tilde{\Rightarrow} \text{MEMBEROF}(U,G).$$

Since the group sentiment is also latent, we include the conceptual inverse to the above rules, which attributes the sentiment of posts by group members to the group's own sentiment. These rules allow this model to collectively infer group sentiment and affiliation:

$$\text{MEMBEROF}(A,G) \tilde{\wedge} \text{POSTED}(A,P) \tilde{\wedge} \text{TAGGED}(P,T) \tilde{\wedge} \text{POSITIVE}(P) \tilde{\Rightarrow} \text{LIKES}(G,T)$$
$$\text{MEMBEROF}(A,G) \tilde{\wedge} \text{POSTED}(A,P) \tilde{\wedge} \text{TAGGED}(P,T) \tilde{\wedge} \text{NEGATIVE}(P) \tilde{\Rightarrow} \text{DISLIKES}(G,T).$$

To enforce consistency in group sentiment, we constrain the truth values of LIKES(G,T) and DISLIKES(G,T) for any group G and tag T to sum to no more than 1.0, which in effect prevents both from being true. We additionally constrain group membership for any individual user to sum to no more than 1.0, such that a user can only fully belong to one group. This last constraint is not always appropriate, depending on the types of groups being considered, but it applies intuitively to the groups we consider in our experiments.

In our experiments, we weight each of these rules uniformly with weight 1.0. In settings where fully-labeled training data is available, we can learn ideal weights for particular data sources. To make predictions with this model, we seed inference with a small set of group affiliations and group sentiment information. The next section describes the application of the model described here to real social media data sets.

3 Experimental Evaluation

This section describes the application of the model from subsection 2.2 to collections of Twitter data. We first provide details on the data sets and data preparation, then analyze the results of PSL group affiliation inference.

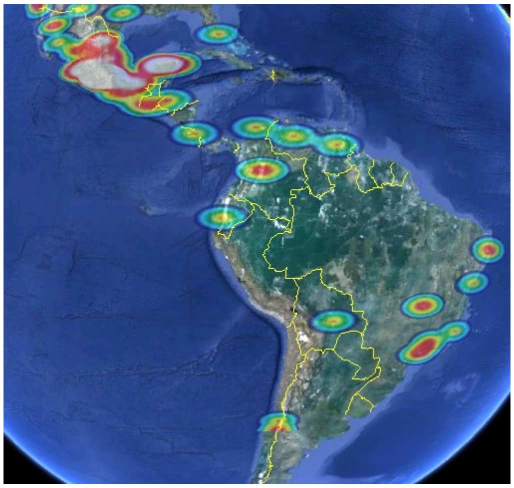
3.1 Data Description

The data we consider is a collection of tweets from the time periods preceding two events: the London 2012 Olympic soccer final match on August 11, 2012 and the Venezuelan presidential election on October 7, 2012. Since the Olympic soccer final was between the Brazil and Mexico teams, we attempt to identify users' affiliations to these teams' fan bases. Similarly, as the main candidates in the Venezuelan election were Hugo Chávez and Henrique Capriles Radonski, we aim to identify supporters of these candidates.

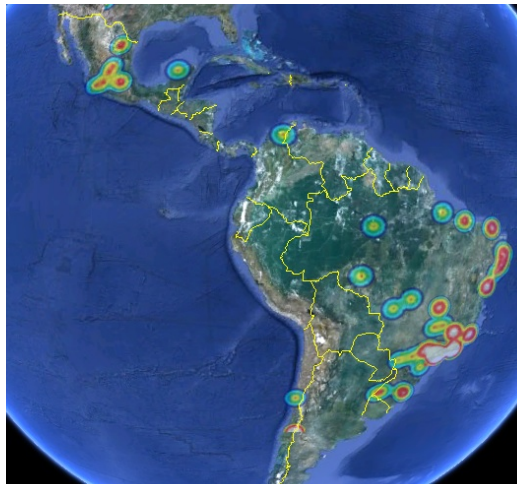
The tweets are filtered to primarily focus on those coming from Latin America, using a variety of indicators, including tweet geotags and Twitter location information. For the Olympic soccer data set, we use tweets from the few hours on August 11 leading up to the game and including the duration of the game. From this period, we have 508,470 total tweets from 316,644 users. In these tweets, 26,457 unique hashtags are used, and 174,380 of the tweets mention other Twitter users. For the Venezuelan election data set, we use the 48 hours (midnight to midnight, Venezuela time) leading up to October 7, from which we have 2,411,472 tweets, 909,933 users, 87,342 unique hashtags, and 1,208,323 mentions. Each tweet is augmented with a sentiment score computed by a third party.[1] We transform this sentiment score with a sigmoid function and map positive values to the truth values of the POSITIVE predicate and negative values to that of the NEGATIVE predicate, counting the truth value of NEGATIVE to be 0.0 when the sentiment is positive, and vice versa.

From an informal overview of the Olympic soccer data, we find there are significantly more tweets in support of the Mexican soccer team than the Brazilian team. The Venezuelan election tweets seem fairly balanced between Chávez and Capriles supporters.

[1] `http://www.datasift.com`



(a) Mexico Group Heat Map



(b) Brazil Group Heat Map

Figure 1: Heat maps indicating the concentration of geotagged tweets from users predicted by PSL to be in each group. The colored scale ranges from blue (light) to red (medium) to white (heavy). Note the heavy concentration of users classified as affiliated with Mexico in the Mexican region on the left, and the heavy concentration of classified Brazil members in the major urban area around Rio de Janeiro on the right.

3.2 Inference Results

For both events, we seed PSL inference with some heuristically labeled instances. To seed soccer fans, we assign users with the strings "mexico", "brasil", or "brazil" in their usernames to be members of their respective namesake groups. This bootstrapped group strategy assigns 174 users to Mexico and 255 users to Brazil. We also seed the tag preferences by asserting that the Mexico group likes the hashtag `#vamosmexico` and that the Brazil group likes the hashtags `#vaibrasil` and `#vamosbrasil`, all with truth value 1.0. We choose these hashtags to seed because they are relatively high precision; a user is unlikely to use these hashtags unless they are expressing their allegiance to the implied fan base.

From this seed information, PSL infers group memberships for 2,913 total users, including the initial seed set. This is a small portion of the total set of users in our data, because the PSL program finds insufficient evidence to assign the remaining users to groups. Adjusting parameters–e.g., priors on the open predicates and the coefficients on the sentiment squashing function–increase the number of inferences. In general, this is expected behavior since the amount of relevant information in social media tends to be sparse. Many tweets are neutral in sentiment, and, for this problem setup, many users may be indifferent and have no allegiance to either group.

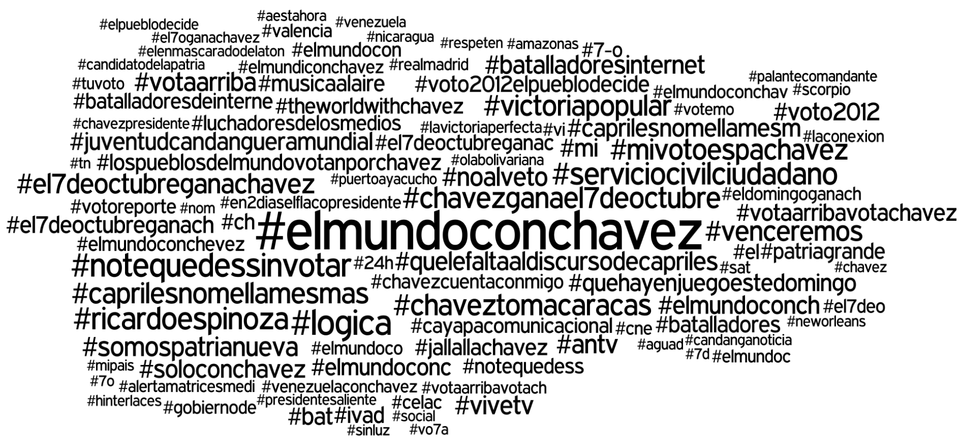
Among the inferred group memberships, a number of users have geotags associated with their tweets. We plot a heat map of the inferred member locations in Figure 1. While the intersection of geotagged users and inferred memberships is relatively small, there is a visible correlation between geographic location within each team's home country and the group affiliations. Specifically, these heat maps are computed by placing a Gaussian bump around the geolocation of any user with MEMBEROF truth value greater than 0.8. Note the high concentration of tweets around major urban areas, such as Rio de Janeiro on the southeast coast of Brazil.

The inferred tag preferences also seem consistent with the desired group identities. Figure 2 contains word clouds[2] of hashtags, with the size of each hashtag weighted by the truth value of the LIKES predicate. A noticeable number of hashtags related to the Mexico soccer victory are linked to the Mexico group. While many Brazil-related hashtags appear in the results from the Brazil group, surprisingly few are related to the Olympic soccer event. This may be a side-effect of the skew

[2]Our word clouds are created using `http://www.wordle.net`.

(a) Mexico Group Preferred Hashtags

(b) Brazil Group Preferred Hashtags



(c) Chávez Supporter Preferred Hashtags

(d) Capriles Supporter Preferred Hashtags

Figure 2: Hashtag clouds for predicted LIKES predicate. The font size is scaled according to the soft truth value of the inferred, latent LIKES predicate.

we notice in the data, or it may be an actual indication of Brazilian Twitter usage during this time period.

For the Venezuelan election data set, we similarly seed the PSL program with hashtag preferences. We identify the hashtags corresponding to popular campaign slogans of each candidate: "#hayuncamino" ("a path forward") for Capriles, and "#elmundoconchavez" ("the world is with Chávez") for Chávez. Again, these hashtags have fairly high precision, compared to, e.g., #chavez and #capriles, which seem to be used by supporters of both candidates, as well as neutral observers and news media. We do not initialize the PSL program with any users heuristically assigned to groups, since there is not as obvious a heuristic in this case (e.g., both Chávez and Capriles are fairly common last names, and would be ineffective heuristics).

The PSL program infers group memberships for 27,713 total users, a much larger number than from the soccer experiment, in part because we use a larger time window for this experiment. The top 100 hashtag preferences inferred by the program for each candidate are displayed in word clouds in Figure 2. The PSL program infers a number of interesting, related hashtags for each candidate, such as #chaveztomacaracas ("Chávez takes Caracas") for Chávez and #chaveztequeda1dia (roughly, "Chávez has one day left") and #unavenezueladepaz ("a Venezuela of peace") for Capriles.

4 Discussion and Future Work

In this paper, we present work on modeling users' affiliations with groups using the declarative modeling language *probabilistic soft logic*. This work is motivated by the future goal of jointly modeling intergroup relational logic with lower-level user affiliation reasoning. By separating group-level relational reasoning from user-level affiliation reasoning, we allow complex relational models at the group level to scale much more than if relational reasoning happens at the user-level. In our preliminary experiment, we simultaneously model the uniformity of social sentiment within groups as well as the tendency for social interactions to be within groups. In future work, we will explore the addition of intergroup sentiment, modeling whether groups pairs are adversarial, cooperative, or indifferent to one another. We are also working to add a temporal component to our models, with the eventual goal of using these multi-level models to predict sweeping changes in social sentiment. Finally, we are using unsupervised topic modeling and clustering methods to initialize the groups, where the goal is to remove the need for the human expert knowledge we use to bootstrap the inference.

Acknowledgments This work is supported in part by the National Science Foundation under Grant No. 0937094 and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC00337. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBA, or the U.S. Government.

References

S. Bach, M. Broecheler, L. Getoor, and D. O'Leary. Scaling constrained continuous markov random fields with consensus optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 44–54, New York, NY, USA, 2006. ACM.

M. Broecheler, L. Mihalkova, and L. Getoor. Probabilistic similarity logic. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.

H. Chang. A new perspective on Twitter hashtag use: diffusion of innovation theory. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, ASIS&T '10, pages 85:1–85:4. American Society for Information Science, 2010.

C. Cooley. *Social Organization: A Study of the Larger Mind*. Social Science Classics Series. Transaction Books, 1983.

M. Hogg. Social identity theory. In P. Burke, editor, *Contemporary Social Psychology Theories*, chapter 6. Stanford University Press, 2006.

J. Lin and A. Kolcz. Large-scale machine learning at twitter. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 793–804, New York, NY, USA, 2012. ACM.

Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. Hellerstein. Graphlab: A new parallel framework for machine learning. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, California, July 2010.

Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. Hellerstein. Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud. *PVLDB*, 2012.

H. Tajfel, M. Billig, R. Bundy, and C. Flament. Social categorization and intergroup behaviour. *European Journal of Social Psychology*, 1:149–77, 1971.

M. Turner and A. Pratkanis. Twenty-five years of groupthink theory and research: Lessons from the evaluation of a theory. *Organizational Behavior and Human Decision Processes*, 73(2â“3):105 – 115, 1998.

L. Yang, T. Sun, M. Zhang, and Q. Mei. We know what @you #tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 261–270, New York, NY, USA, 2012. ACM.

E. Zheleva, H. Sharara, and L. Getoor. Co-evolution of social and affiliation networks. In *15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, June 2009.