

Assessment of Effectiveness of Content Models for Approximating Twitter Social Connection Structures

Kuntal Dey^{*†}, Sahil Agrawal[†], Rahul Malviya[†] and Saroj Kaushik[†]

^{*}IBM Research India, New Delhi, India

[†]Indian Institute of Technology, New Delhi, India

Abstract—This paper explores the social quality (goodness) of community structures formed across Twitter users, where social links within the structures are estimated based upon semantic properties of user-generated content (corpus). We examined the overlap of the community structures of the constructed graphs, and followership-based social communities, to find the social goodness of the links constructed. Unigram, bigram and LDA content models were empirically investigated for evaluation of effectiveness, as approximators of underlying social graphs, such that they maintain the *community* social property. Impact of content at varying granularities, for the purpose of predicting links while retaining the social community structures, was investigated. 100 discussion topics, spanning over 10 Twitter events, were used for experiments. The unigram language model performed the best, indicating strong similarity of word usage within deeply connected social communities. This observation agrees with the phenomenon of evolution of word usage behavior, that transform individuals belonging to the same community tending to choose the same words, made by [1], and raises a question on the literature that use, without validation, LDA for content-based social link prediction over other content models. Also, semantically finer-grained content was observed to be more effective compared to coarser-grained content.

I. INTRODUCTION

User generated content on social networks such as Facebook, and microblogs such as Twitter, has become a trending research topic in recent years. Microblogs have been studied from a number of research perspectives, such as information diffusion [2] and spread of ideas [3].

A. Motivation

Several works, such as [4], [5], [6], [7] and [8] address predicting (constructing) social links between pairs of users, from graph attributes. [9] and [10] focus on graph structure and properties. [11] predicts links based upon semantic content. [12] studies language-based conversation modeling of Twitter users.

Statistical count of correctness of predicted links does not reveal any structural or social insight about the prediction. Two cases of predicting links using two independent algorithms could have similar precision and recall, but could form completely different graph (social) structures. This would imply radically different community structures and dynamics.

The goodness of the predicted social links is measured in literature, including [11], using statistical phenomenon such as accuracy, precision and recall, not *social* attributes. [11] selects Latent Dirichlet Allocation (LDA) [13] without exploring

other language models. Also, no work investigates the impact of different user-generated content (corpus) granularities. This necessitates our work.

B. Contributions of Our Work

In the current work, we overlay the community structures formed by two graphs: a content graph and a social friendship graph. Content graph is created by collecting the lifetime content generated (tweets made) by the user, related to a given event. The links between user pairs, are constructed (predicted) based upon the content generation similarity between user pairs, as found by content models, such as unigram, bigram and LDA. We use the popular modularity [14] maximization technique, which is clearly by far the most widely accepted definition of social network communities [15]. We specifically apply the BGLL [16] algorithm implementing the technique, to identify the structural communities implicitly present in the constructed graphs (predicted links). Community overlap of each pair of graphs is quantified using normalized mutual information (NMI) [17]. A higher NMI value indicates a better prediction goodness.

100 different topics, spanning over 10 different Twitter events, were used for experimentation. The hashtags were chosen for popular events, and each hashtag chosen was unique to a event within the dataset. For each topic and each event as a whole, significant community overlaps were found between the two graphs, as indicated by NMI values. Among the unigram, bigram and LDA models, the unigram model was observed to provide the best overlaps. This is revealing: it suggests significant likeness of users within social communities, in terms of word (language) usage. Interestingly, this is in spirit similar to the made by [1], where it is observed that the language usage of social friends evolve and become similar over time (word usage similarity increases). Further, fine-grained topics found by running LDA on the whole of user content, provided a better approximation (prediction) of the social graph structure, compared to the coarser-grained events represented by hashtags. The knowledge thus obtained can be used in applications such as social information flow modeling and marketing.

II. RELATED WORK

Link prediction on social networks has been an area of long standing research. [7] carried out a comprehensive study of

different link prediction techniques in a social network setting, including methods such as graph distance, Adamic-Adar method [4], Jaccard’s coefficient [8], rooted Pagerank [7], Katz [6] and SimRank [5]. However, this study, and the subsequent ones in this school of research such as [10] and [9], focuses on graph structure and properties, and does not consider content semantics. Subsequently, [11] attempted to study the impact of communication semantics in predicting social links, and used Twitter as their platform for the study. This study uses LDA [13] for predicting pairwise links; however, it neither investigates the relative impact of different language in such prediction, nor does it delve deeper to investigate the social properties of the predicted links, which is essential if one were indeed aiming to predict a *social* network.

Researchers have attempted to investigate the flow of information cascades on Twitter, as well as propagation of influence along the underlying social connection graph. [3] predicts the spread of user-generated ideas on Twitter. [18] proposes a multi-class classification model to identify popular messages on Twitter, by predicting retweet quantities, from TF-IDF (term frequency and inverted document frequency) and LDA, along with social properties of users. [19] models the flow of influence along social connections on Twitter, and makes the surprising observation that in spite of URLs rated interesting and content by influential users spreading more than average, predictions of which particular user or URL will generate large cascades are relatively unreliable. Other studies, such as [20], [21], [22], [23], [24] and [25] provide significant insights into flow of information and influence, along social edges, over Twitter user interactions.

In order to find social communities for exploring our problem space, the current work makes use the community finding literature. The most prominent class of implicit communities formed based upon graph structures is modularity-based communities. Originally proposed by [14], a fast approximation algorithm is used, BGLL [16], to compute max-modularity communities. In order to derive modularity values, this body of work initially computes the differences of actual and expected (probabilistic) value of a given pair of vertices to have an edge, and subsequently aggregates the above over all possible pairs to maximize the value of modularity. Cross-entropy [26] is also used, which is based upon Kullback-Leibler (K-L) divergence [27], and normalized mutual information [17], as part of our computation methodology.

Using a combination of social links and user-generated content has been explored in the literature, from different angles. [28] attempt to combine the strength of links with similarity of content between each pair of graph vertices (social network users), to augment baseline social link based graphs, and discover communities on such graphs. [29] combine the topological structure of a network with the content information present in the network, and thereby model the community structure as a consequence of interaction amongst the participating nodes (social network users). [30] also combine the graph node attributes with the graph edge structures for community discovery. [31] also consider the overlaps between communities using the concept of the intersection graph, for community discovery. [32] observe that joint modeling of links

and content significantly improves link prediction performance on Twitter subgraphs.

Clearly, there is significant background prior art that exists in related areas. Some semantic link predictions exist, and some literature also attempt to integrate content with link for community discovery. However, no direct in-depth investigation exists that attempts to benchmark or compare the goodness of different language models, in predicting or approximating microblog connections while capturing the essential community structures. Further, no question has ever been raised in the literature on the quality of social structure prediction, or even simple link prediction, with respect to topic identification granularities, with respect to any content or semantic attribute. We attempt to answer these fundamental questions that have not been investigated in the literature. This makes our work a first of its kind.

III. PROPOSED APPROACH

This section explores the problem settings, and the machinery developed to obtain insights under these settings.

A. Problem Settings

The core objective of the current work is to approximate (predict) the microblog connection graph, from user-generated content. In the process, the following have been proposed.

- Quantifying the *social* goodness of microblog friendship graphs, approximated (predicted) using user-generated content.
- Investigating the goodness of different language models for the approximation.
- Studying the impact of granularity of selecting content, to predict the graph structures.

B. Solution Framework

Our solution framework overlays the structures (communities) formed by two graphs: a graph created by connecting user pairs based upon user-generated content, and a social friendship graph given as ground-truth. The overlaps of the sets of communities C_L in the content graph and C_S in the social graph are quantified, to obtain insights. Figure 1 depicts the architecture of our system.

A content graph is first constructed. The topics present in the document are detected, using LDA. For each user and each topic, all the tweets that the user makes in her lifetime for the topic are collected. The set union of these tweets are used to create a document for the user for that topic, as described in Section IV-A.

In order to construct semantic edges between user pairs, different language models, such as unigram, bigram and LDA (topic model), are applied. The techniques to construct these edges (links) are different in case of LDA, from unigram and bigram: the edge construction methodology is explained in Section IV-B. Effectively, these edges are associated with a confidence score, that represents the similarity between user pairs. The predicted links (constructed edges) are now further processed to extract structural social communities,

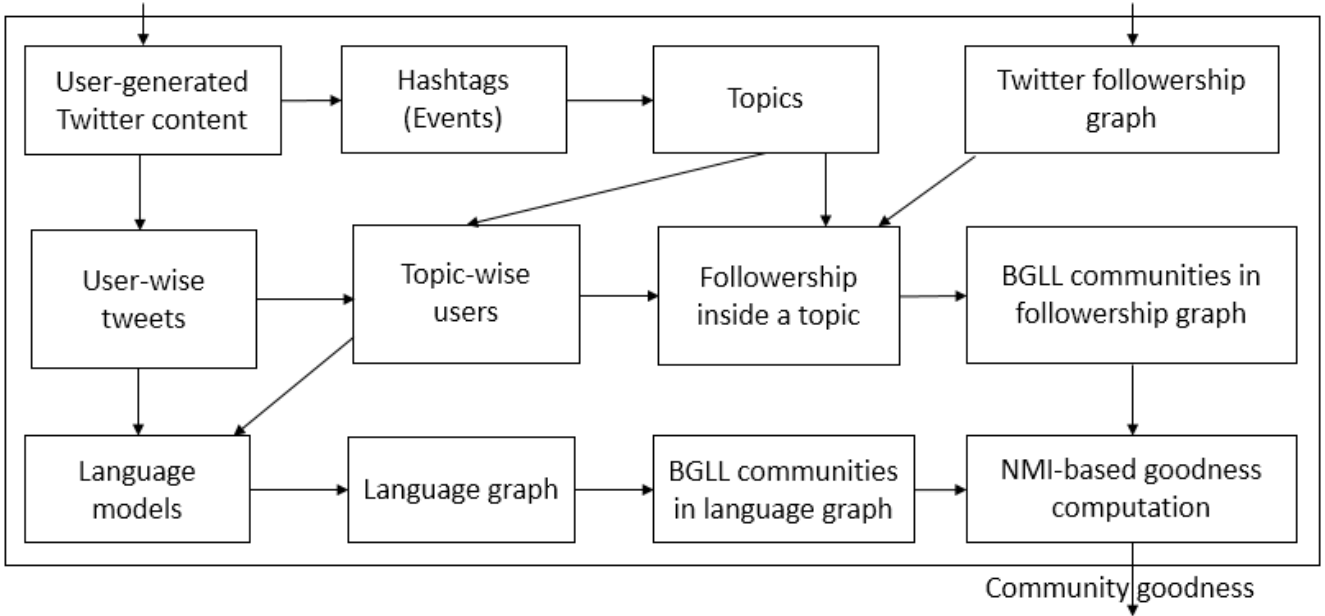


Fig. 1. The architecture of our framework

which is used to evaluate the goodness of the prediction, as detailed in Section IV-C. For this, the spectral modularity maximization techniques of Newman [14], implemented by BGLL [16], is used. The ground-truth communities, using known followership (friendship) data present in the social graph, are discovered. The overlap of these sets of communities are subsequently computed, quantifying using normalized mutual information (NMI). Higher NMI values indicate better overlap, and hence, better prediction (approximation) quality. The overview of the approached solution framework shown in Algorithm 1.

IV. IMPLEMENTATION

The key steps of the implementation process are described below.

A. Document Creation and Topic Extraction

Documents are generated for each user, as well as the topics present in the content under consideration are extracted. These topics, and the documents, are subsequently used to create the content graph, by drawing language-based edges between user pairs.

For this, the content generated by each user on the microblog is collected, that contains the hashtag H_e . H_e uniquely identifies event e , from a set of events E . Since each tweet is small in length, within 140 characters (typically 14-18 words), the set union of all tweets t_{eu_i} made by user u_i , that contain the hashtag H_e , is computed. This forms document d_{eu_i} , representing the lifetime participation of user u_i in event e , given by Equation 1. Let T_e denote the complete set of tweets belonging to event e .

$$d_{eu_i} = \bigcup_{(t_{eu_i} \in T_e)} \{t_{eu_i}\} \quad (1)$$

Having constructed the individual documents using Equation 1, the topics that are present within the events are extracted. The overall (broader) document for a given event is extracted, comprising of all tweets having the current event hashtag, as shown in Equation 2.

$$D_e = \bigcup_i \{d_{ei}\} \quad (2)$$

Subsequently, the LDA model is applied, to find the intended topics. A well-recognized tool, MALLET [33], is used to detect topics $C_e = \{c_{ek}\}$ for a given event e , for our experiments. These topics c_{ek} , as well as the per-user documents d_{ei} , are subsequently used for creating the content graph, and thereby for modeling the weighted semantic links between pairs of users.

B. Content Graph Creation

Creating the content graph involves identifying content creation similarity among user pairs, measured by a chosen language model, and thereby identifying links between the pair. Two classes of language models are selected: (a) the unigram and bigram models from the n -gram language model class, and (b) the LDA model from the probabilistic semantic model class.

Edge creation in n -gram models

The n -gram language model l_i is computed for each user u_i that have participated in any of the events, and any topic of the event. It should be noted that the language model is independent of any event or topic that the user participates in. For each given event e , edges between each pair of user participating in that event are created. Cross-entropy [26] being inherently asymmetric, to create an edge between a given

Algorithm 1 OVERVIEW OF OUR APPROACH

Input 1. User-generated Twitter content

Input 2. Twitter followership graph

```

1: select a set of hashtags (events) from Twitter content
2: assign a user  $u_i$  to all events (hashtags) that she participates in
3: for each event  $e$  do do
4:   identify topics  $c_{e,k}$  within event  $e$ , using LDA
5:   for each user  $u_i$  in event  $e$  do do
6:     find event-level goodness value as the NMI of language and social graph communities
7:     construct document  $d_i$  as collection of all tweets  $t_{e,u_i}$  of  $u_i$ 
8:     for each topic  $c_{e,k}$  do do
9:       find users  $u_{i,e,k}$  participating in topic  $c_{e,k}$ 
10:      find language models of these users, using topics and tweets
11:      create language graph by drawing edges between user-pairs
12:      assign language graph edge weight inversely proportional to user-pair cross entropy
13:      identify BGLL communities in the language graph
14:      create topic-level social graph of user-pairs, using Twitter followership graph
15:      identify BGLL communities in the topic-level social graph
16:      find topic-level goodness value by finding NMI of language and social graph communities
17:     end for
18:   end for
19: end for

```

Output: Set of community goodness scores

pair of users u_{e,i_1} and u_{e,i_2} , we take the average of cross-entropy of the language model l_1 of user u_{e,u_1} with respect to the other user's event-level document d_{e,u_2} , and that of l_2 w.r.t. d_{e,u_1} . Weights are assigned to the edges as the difference of the maximum value of the cross-entropy and the value of the cross-entropy of the current edge (since high cross-entropy denotes low similarity). Thresholds are subsequently applied, to finally retain or discard the edge created hereby, by selecting the top $k\%$ edges for experimentation.

Apart from constructing the edges for the event-level graphs, constructing edges for the topic-level graphs is also necessary, for the topics that were found by the LDA model within each event. For each document d_{e,u_i} , we find the probability $p_{e,d_{e,u_i},c_{e,k}}$ of the document to belong to a topic $c_{e,k}$. If there are K topics for the event e , then it is assumed that a user is associated with the topic as long as $p_{e,d_{e,u_i},c_{e,k}} > \frac{1}{K}$. The intuition behind this is simple: it retains all the topics in which the user's participation was higher than the expected random participation, denoting user interest in the topic. This will create a subset of users u_i , that participate in event e . The cross-entropy based edge-creation process is repeated at the topic level also. For each topic, weight inversely proportional to the cross-entropy values is assigned. This creates one graph per topic, within a given event.

Threshold-based retention of the event-level or topic-level edges, generated by the above, is subsequently applied. For experimentation, instead of considering any absolute threshold (it is scientifically difficult to establish a *good* threshold), the goodness of our system is tested by retaining a certain percentage of the best edges (high values of weights), and discarding the rest. This is useful because it provides an

intuition into optimally selecting (predicting) edges that can be easily identified, rather than all edges, for our current purposes.

It should be noted that, if $H(P)$ denotes the marginal entropy of a probability distribution P , the cross-entropy of a probability distribution Q w.r.t. P is given as:

$$H(P, Q) = H(P) + D_{KL}(P||Q) \quad (3)$$

Here, $D_{KL}(P||Q)$ is the K-L divergence [27] of Q , a probability distribution, w.r.t. P , another probability distribution. Further, for a discrete distribution such as ours, K-L divergence is given by

$$D_{KL}(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \quad (4)$$

The n-gram model based content edge creation process is applied for $n = 1$ (unigrams) and $n = 2$ (bigrams). The n-gram probability distribution Q , and the document as P , are used to compute cross-entropy.

Edge creation in the LDA model

In order to create the edges in the LDA model, the probability $q_{e,d_{e,u_i},c_{e,k}}$ is computed for each user u_i , for participating in each topic, by assigning the user's document d_{e,u_i} a probability to belong to each topic $c_{e,k}$. If there are K topics for the event e , then it is assumed that a user is associated with the topic as long as $q_{e,d_{e,u_i},c_{e,k}} > \frac{1}{K}$. This creates a subset of users of the event, to participate in the topic. An event-level edge between a given user pair, say u_1 and u_2 , with weight computed as a sum-of-product of the probabilities (Equation 5), is created.

$$w_{u_1, u_2}^{LDA} = \sum_k (q_{e, d_e, u_1, c_{ek}} * q_{e, d_e, u_2, c_{ek}}) \quad (5)$$

In case of a topic model, a similar process is applied on the subset of users participating in the topic. Subsequently, a threshold-based edge retention process is carried out, in a manner similar to what was done in the n-gram model. This completes the process of constructing the content graph (effectively, prediction of the links). The three sets of graphs derived from the three language/topic models, namely unigram, bigram and LDA, and each constructed at two granularities, namely event and topic levels, are now assessed for structural goodness.

C. Measuring Structural Goodness

Most of the related literature, including the LDA-based edge prediction by [11], has measured the goodness of their work, using prediction accuracies and error rates. On the contrary, the current work aims to measure the *social* goodness of the predicted links. Therefore, the overlap of the communities (social structures) formed by the predicted links (semantic graph), with the ground truth (social graph), is quantified. Thus, the prediction is effectively evaluated at a structure level, which has much deeper social semantics compared to individual links.

This is done in two stages. First, the modularity-based structural communities in the content graph, as well as the social friendship graph, are independently detected. Subsequently, the overlap two community sets is measured, by computing the NMI within each of these two sets. By definition [17], NMI values range between 0 and 1, and a higher NMI value indicates a higher overlap of the two sets of communities.

Finding modularity communities

The concept of modularity [14] is used to discover implicit structural communities. Modularity technique aims to partition a given graph into non-overlapping components, maximizing the proportion of connections between pairs of vertices belonging to the same component, to that of pairs of vertices belonging to two different components. To compute modularity, two quantities are considered for a given pair of components. (a) The ground truth, which is deterministic about whether a given pair of vertices are connected by an edge, or not. (b) The probabilistic expectation that a given pair of vertices is connected, given the total degree and total number of vertices of the entire graph. The partitioning is carried out by computing the differences from the former with the latter, and aggregating over the arrangements.

Newman's spectral method for computing modularity is formulated as:

$$Q = \frac{1}{4m} \sum_{i,j} (A_{ij} - \frac{k_i k_j}{2m}) s_i s_j \quad (6)$$

Here Q denotes the modularity, A_{ij} are the adjacency matrix elements (edge) between vertices i and j , $k_i k_j / 2m$ is the expected number of edges between vertices i and j when

placed at random, $1/4m$ is a conventional factor and s_i and s_j are components (communities) that vertices i and j belong to. Newman's method is computationally expensive, taking $O((m+n)n)$ time, where n is the number of vertices and m is the number of edges in the graph. Instead, the BGLL [16] algorithm is used, since it provides a fast implementation of modularity computation.

Computing NMI

Mutual information is computed as

$$I(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right)$$

where $I(X, Y)$ is the mutual information of X and Y . The normalized mutual information (NMI) [17] across these two sets of communities is computed as

$$C_{XY} = \frac{I(X, Y)}{H(Y)}$$

and

$$C_{YX} = \frac{I(X, Y)}{H(X)}$$

respectively, where $H(X)$ and $H(Y)$ denote the marginal entropies of X and Y respectively.

V. EXPERIMENTS

Experiments were carried out on data from 10 Twitter events, having unique hashtags. Further, 10 LDA-based topics were identified within each event using MALLET [33]. We thus validate our hypotheses experimentally, over 10 events and $10 * 10 = 100$ topics. The tools used were: MALLET, to find topics and find the LDA-based content model of users, and statistical language modeling toolkit by [34] for unigram and bigram models, and cross-entropy.

A. Data Description

Twitter data was collected over 10 different events, where each event was identified by a unique hashtag. The following facets of the data were collected: (a) For each hashtag, all data (content) that was generated. (b) All tweets (content) that each user, who ever posted with a given hashtag, in their lifetime on Twitter (limited by Twitter max of 3,200). (c) Followership graph of each of these users. Implicit reciprocity in the followership graph is assumed. For the sake of brevity, we present the experimental results for 60 topics, spanning over 6 randomly chosen events from the 10. Table I presents basic statistics of these events.

B. Evaluating Content Models

Content models, namely unigram and bigram language models and LDA topic model, were evaluated at this stage. Figure 2 shows the variation of NMI for each content model, at different threshold levels, across multiple events. The results obtained with bigrams were below par, hence the experiments presented were conducted with unigram and LDA. The threshold levels

TABLE I
BASIC STATISTICS OF EVENT DATASETS. EDGES ARE BASED ON FOLLOWERSHIP, IGNORING DIRECTIONS.

Event Hashtag	Num Nodes	Num Edges	Num Tweets
Billboards	327	778	7,579
Cesar	1,005	6,013	5,273
Coachella	453	604	8,876
Elections	473	1,575	8,815
Junos	624	3,366	4,334
Ted	631	1,243	7,184

have been chosen to retain the top $K\%$ of the edges (ranked by weights).

In our experiments, the unigram models performed better than LDA consistently across all the datasets. This is clear from the NMI values, which are higher for unigram, compared to LDA. Plots for 6 events are shown in Figure 3; however, these trends were consistently observed for all the events and topics explored. This indicates significant similarity of word usage within social communities, which is well-captured by unigram, but given theme (topic) similarity would create confusions in case of LDA. Manual inspection of ground-truth friendship edges, specifically predicted by unigram but not by LDA, confirm these trends.

C. Evaluating Granularities

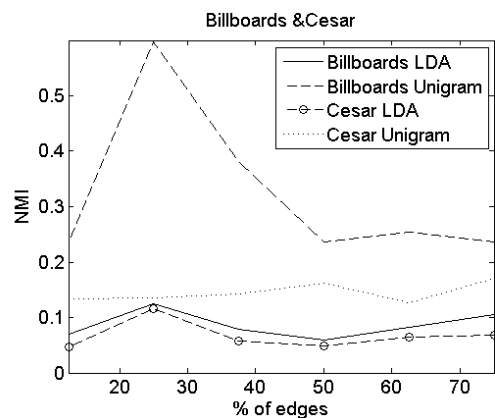
The impact of granularity of the chosen corpus, namely the coarser-grained event and finer-grained topic level granularity, on NMI values, was evaluated over different thresholds. Figure 3 demonstrates the trends. Finer-grained topics consistently yielded higher NMI values, compared to coarser-level events. This shows the effectiveness of selecting finer-grained topic-level tweets as corpus to predict social structures, over coarser-grained event-level tweets.

D. Observing Community Structures

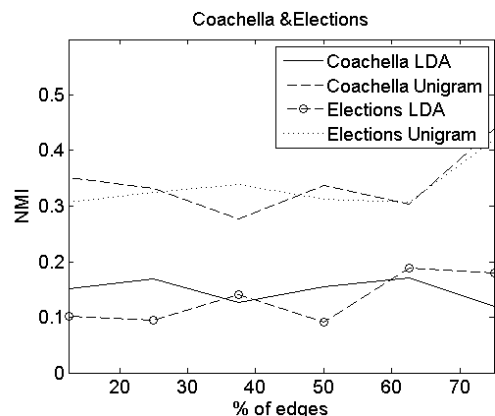
Our experiments revealed significant NMI values on the content based graph communities, with respect to the explicit friendship based social communities. Since friendship connections are ground truth, the NMI values indicate that the content models have been able to retain much of the communities, a core social property, in the predicted links. This demonstrates the effectiveness of content models, in predicting “good” social links from user-generated text in Twitter, with the unigram model being the most effective (validated by retaining the community structures better than the other models, for all the datasets). Finally, finer-grained topics were observed to be better approximators of social communities, compared to coarser-grained events.

VI. DISCUSSION

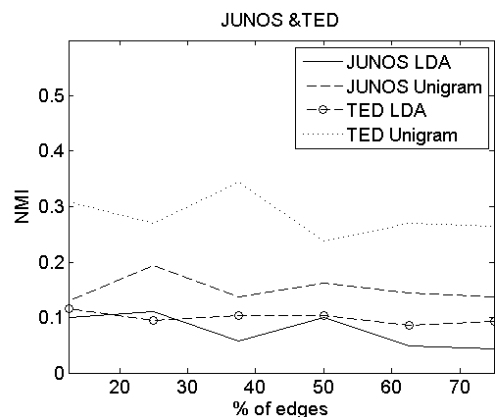
The current study elicits a surprising insight: in spite of being inherently a much simpler model compared to LDA, the unigram model was seen to be more effective in capturing the links that are better in attaining a deep and complex social property - social communities. This can be attributed to word



(a) Events: Billboards, Cesar



(b) Events: Coachella, Elections



(c) Events: JUNOS, TED

Fig. 2. Effectiveness of content models as approximators of social structures (communities). 2 hashtags covered per plot.

usage behavior similarity within social communities (exact same words getting used within communities), favoring the unigram model, leading to spurious edges in LDA.

The authors have noted that in recent literature, the language usage behavior of individuals has been shown to evolve and become similar to other individuals belonging to the same communities, over time [1]. [1] exemplify with *aroma* and *smell*: they observe that communities built upon the keyword *beer*, at one point of time, tended to use the word *aroma* together, which over time evolved into *smell* (or, S in short).

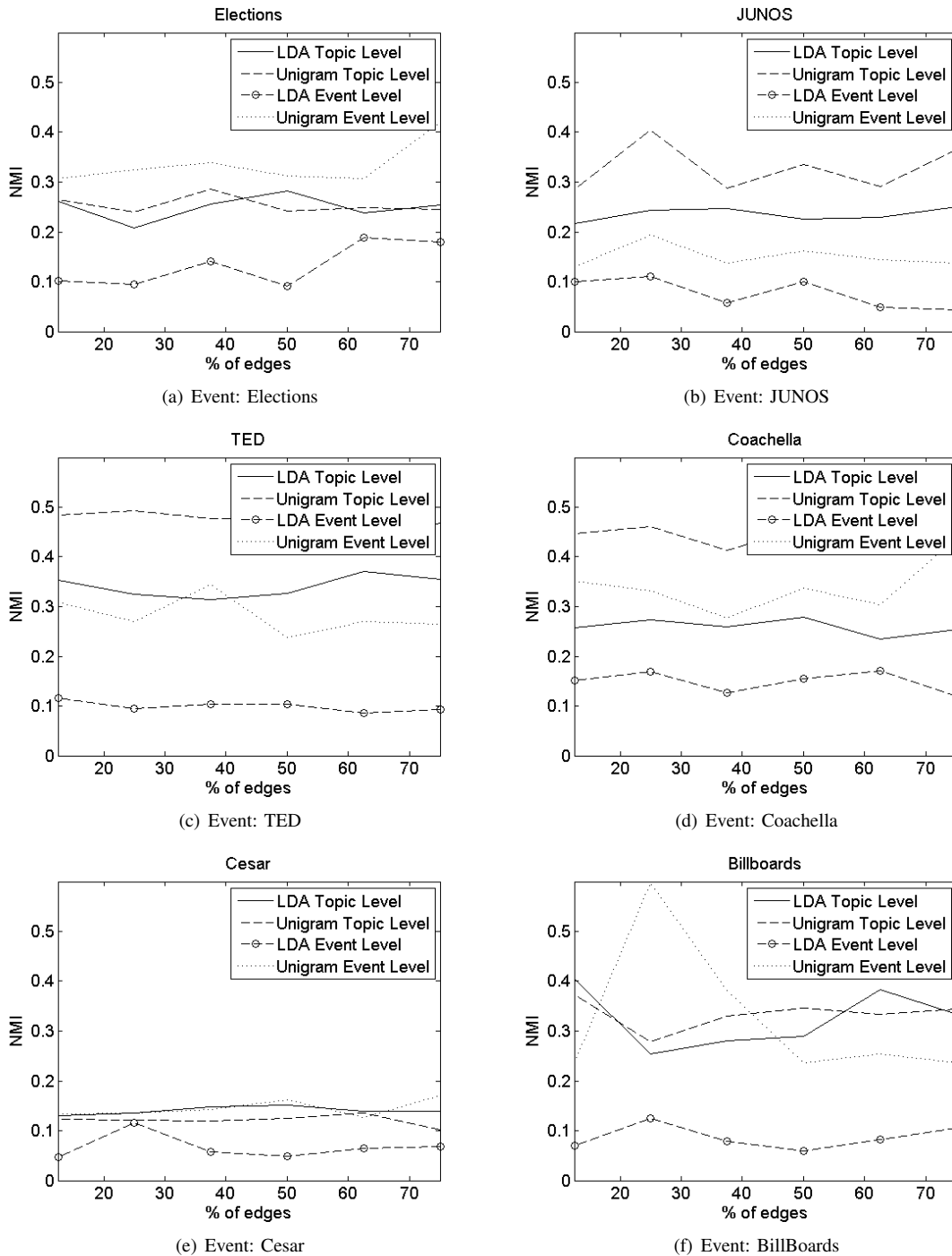


Fig. 3. Impact of corpus granularity selection on NMI, at topic (fine) vs. event (coarse) levels.

However, the phenomenon of using the same word (effectively, *unigram similarity*) was observed among the wide span of connected individuals belonging to the communities, not just at the level of individuals. Note that, this observation is strongly in alignment with our observations of unigram outperforming LDA: LDA would capture *smell* and *aroma* to be similar, and use that as a predicting feature of a social connection, while unigram will not. The unigram match outperforming other content models is likely to be a reflection of this very phenomenon. In other words, while LDA gives a probabilistic

distribution of language usage for individuals, unigrams, the specific words used by individuals, match with each other as individuals belonging to the same community pick up word usage behavior of others and tend to start using similar words. This is an interesting observation about the deeper language usage behavior of individuals belonging to similar communities.

Thus, while much of the current literature adapt LDA for predicting content-based social links, we are the first ones in the literature to raise a question regarding whether, in spite of

its richness, LDA is at all the most effective content model, and surprisingly observe on the contrary. Further, since our work captures the tweets made by each user under consideration for their entire lifetime on Twitter that is made publicly available, our LDA model is trained on as much Twitter data that one can access, and what every other content based link construction is made from in the literature, that uses the public Twitter APIs.

VII. CONCLUSIONS

In the current work, a framework was created to study the effectiveness of language models in approximating or predicting microblog connection structures. Hashtags were used to identify coarse-grained events. LDA-based fine-grained topics were found within each event. The participation of users were found, in each topic within each event, from a given set of events, based upon user generated content. Using the language usage similarities and topic similarities of all user pairs, a content-based user graph was created, spanning participants of the event-related discussions. Unigram, bigram and LDA were used as the underlying models. For each event, the overlap of the language graph structure, with the ground-truth social graph structure, was quantified at different content granularities such as event-level and topic-level, using NMI. Experiments were conducted with 100 topics, spanning over 10 Twitter events, for empirically proving our proposition. The results consistently demonstrate the goodness of the approximation, at different granularities, with higher NMI values emanating from more fine-grained topics. The unigram model was consistently found to be most effective, in all the cases. This indicates a strong similarity of word usage behavior of users within deeply connected social communities. Some applications of the current work would be in the academic area of information flow modeling, as well as practical field of social marketing based applications.

REFERENCES

- [1] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts, "No country for old members: User lifecycle and linguistic change in online communities," in *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013, pp. 307–318.
- [2] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *WWW*. ACM, 2010, pp. 591–600.
- [3] O. Tsur and A. Rappoport, "What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities," in *WSDM*. ACM, 2012, pp. 643–652.
- [4] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social networks*, vol. 25, no. 3, pp. 211–230, 2003.
- [5] G. Jeh and J. Widom, "Simrank: a measure of structural-context similarity," in *KDD*. ACM, 2002, pp. 538–543.
- [6] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.
- [7] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.
- [8] G. Salton and M. J. McGill, "Introduction to modern information retrieval," 1986.
- [9] Y. Dong, J. Tang, S. Wu, J. Tian, N. V. Chawla, J. Rao, and H. Cao, "Link prediction and recommendation across heterogeneous social networks," in *ICDM*. IEEE, 2012, pp. 181–190.
- [10] D. Yin, L. Hong, and B. D. Davison, "Structural link analysis and prediction in microblogs," in *CIKM*. ACM, 2011, pp. 1163–1168.
- [11] K. Puniyani, J. Eisenstein, S. Cohen, and E. P. Xing, "Social links from latent topics in microblogs," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*. Association for Computational Linguistics, 2010, pp. 19–20.
- [12] A. Ritter, C. Cherry, and B. Dolan, "Unsupervised modeling of twitter conversations," 2010.
- [13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.
- [14] M. E. Newman, "Modularity and community structure in networks," *PNAS*, vol. 103, no. 23, pp. 8577–8582, 2006.
- [15] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri, "Extending the definition of modularity to directed graphs with overlapping communities," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, no. 03, p. P03024, 2009.
- [16] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.
- [17] C. H. Coombs, R. M. Dawes, and A. Tversky, "Mathematical psychology: an elementary introduction." 1970.
- [18] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in *Proceedings of the 20th international conference companion on World wide web*. ACM, 2011, pp. 57–58.
- [19] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in *WSDM*. ACM, 2011, pp. 65–74.
- [20] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *WWW*. ACM, 2012, pp. 519–528.
- [21] S. A. Myers, C. Zhu, and J. Leskovec, "Information diffusion and external influence in networks," in *KDD*. ACM, 2012, pp. 33–41.
- [22] K. Narang, S. Nagar, S. Mehta, L. V. Subramaniam, and K. Dey, "Discovery and analysis of evolving topical social discussions on unstructured microblogs," in *Advances in Information Retrieval*. Springer, 2013, pp. 545–556.
- [23] D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter," in *WWW*. ACM, 2011, pp. 695–704.
- [24] J. Yang and S. Counts, "Predicting the speed, scale, and range of information diffusion in twitter," *ICWSM*, vol. 10, pp. 355–358, 2010.
- [25] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," in *ICDM*. IEEE, 2010, pp. 599–608.
- [26] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.
- [27] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [28] Y. Ruan, D. Fuhry, and S. Parthasarathy, "Efficient community detection in large networks using content and links," in *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013, pp. 1089–1098.
- [29] L. Liu, L. Xu, Z. Wang, and E. Chen, "Community detection based on structure and content: A content propagation perspective."
- [30] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *Data mining (ICDM), 2013 IEEE 13th international conference on*. IEEE, 2013, pp. 1151–1156.
- [31] T. Kuramochi, N. Okada, K. Tanikawa, Y. Hijikata, and S. Nishida, "Community extracting using intersection graph and content analysis in complex network," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, vol. 1. IEEE, 2012, pp. 222–229.
- [32] N. Natarajan, P. Sen, and V. Chaoji, "Community detection in content-sharing social networks," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2013, pp. 82–89.
- [33] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002.
- [34] R. Rosenfeld and P. Clarkson, "Statistical language modeling using the cmu-cambridge toolkit," 1997.