Mining Divergent Opinion Trust Networks through Latent Dirichlet Allocation

Nima Dokoohaki
Software and Computer Systems (SCS)
Royal Institute of Technology (KTH)
Forum 120, 16440-Kista, Sweden
Email: nimad@kth.se

Mihhail Matskin
Software and Computer Systems (SCS)
Royal Institute of Technology (KTH)
Forum 120, 16440-Kista, Sweden
Email: misha@kth.se

***Abstract*—While the focus of trust research has been mainly on defining and modeling various notions of social trust, less attention has been given to modeling opinion trust. When speaking of social trust mainly homophily (similarity) has been the most successful metric for learning trustworthy links, specially in social web applications such as collaborative filtering recommendation systems. While pure homophily such as Pearson coefficient correlation and its variations, have been favorable to finding taste distances between individuals based on their rated items, they are not necessarily useful in finding opinion distances between individuals discussing a trending topic, e.g. Arab spring. At the same time text mining techniques,such as vector-based techniques, are not capable of capturing important factors such as saliency or polarity which are possible with topical models for detecting, analyzing and suggesting aspects of people mentioning those tags or topics. Thus, in this paper we are proposing to model opinion distances using probabilistic information divergence as a metric for measuring the distances between people's opinion contributing to a discussion in a social network. To acquire feature sets from topics discussed in a discussion we use a very successful topic modeling technique, namely Latent Dirichlet Allocation (LDA). We use the distributions resulting to model topics for generating social networks of group and individual users. Using a Twitter dataset we show that learned graphs exhibit properties of real-world like networks.**

I. Introduction

Social Web is collectively perceived as an aggregate notion of users identities (profiles) and belongings (contributed or customized content) linked across multiple networks [1]. As user generated content remains the most significant proportion of users belongings across social sites on-line today, information overload poses a challenge for service providers trying to leverage this content to users benefit and value through customization or personalization functions. A significant amount of this contribution is basically natural language spoken content such as tweets, in the case of Twitter or status updates, in the case of Facebook.

While leaving size and format of this content aside, analyzing it is a challenge for both content providers and consumers. Thus an increasing number of works are focused on proposing analyzing natural language content from social services for the benefit of users and services [2]. While computational techniques are being proposed for analyzing spoken text on social networks, opinion mining techniques are increasingly attractive to analyze networks of users informing other like minded ones across the social sphere [3].

Topic modeling mechanisms [4] are increasingly attractive, due to their success in mining diverse opinions from e-commerce web, specially consumer review sites [5]. Thus an increasing number of researchers are proposing for their adoption within social web domain. Due to their probabilistic nature, it's possible to build social networks out of resulting mixture of topics and their associated distributions [6]. While these networks have been limited to associating terms and authors [7], or communities and tags [8], modeling trust networks have not been of significant attention. Moreover, due to their probabilistic learning approach proposing divergence metrics as distances between nodes on the network is of novelty. Since it can model diverse relations among users, while using the topic modeling can allow for aspects like saliency, relevancy and even polarity to be measured amongst networked opinions.

Thus in this work we are proposing for a topic modeling framework, within which a trend corpora can be mined and by using a Latent Dirichlet Allocation (LDA) technique, both collective, and individual models can be defined. Resulting models are eventually used to generate social networks which reflect divergences of collective and individual opinions. This is followed by an experiment on a Twitter dataset to justify that resulting networks own properties of real-world social networks, both in structure and in content. While these weighted graphs are focus of social network analytics within this manuscript, we plan to leverage them for tasks of filtering and recommendation.

Finally, this paper is segmented into the following manner: first a comprehensive background is presented. This is followed by framework description, within which each step is outlined and described in detail. Then experiment is outlined and followed by conclusion remarks and future works.

II. Related Work

We divide this part into two overlapping sections, one outlining topic mining techniques for social network analysis which is followed by a section focusing on using topic modeling for trust modeling and mining.

A. Social Network Analysis and Topic Models

Topic models are of great importance in opinion mining and summarization literature. So before focusing on topic models and their importance with respect to social network analysis and mining a brief introduction to opinion mining deems necessary.

An opinion summary encompasses any study that attempts to generate a concise and digestible summary of a large number of opinions. A modern opinion summary boils down to a structured digest that provides a well-organized breakdown by aspects/topics, various formats of textual summaries and temporal visualization. In a recent study, Kim, et al.[9] give a multi-perspective classification of approaches to opinion summarization and integration. Chen and Zimbra report on several applications of opinion mining in various contexts [10].

Opinion summarization is more and more attractive to social networking analysts. Opinion summarization and mining encompasses an increasingly large range of applications on social web but a few are most common among others in recent literature: personalized recommendation systems i.e. tag-based or collaborative filtering-based [11], sentiment analysis for instance in Twitter verse [12], [13], [3]. Among such techniques under focus, topic models have been very successful. Topic models are generative probabilistic models which utilize vocabulary distillations to spot topics within text corpora. Most widely utilized topic modeling techniques include Probabilistic Latent Semantic Analysis (PLSA) [14] and Latent Dirichlet Analysis (LDA) [4]. There exists applications of LDA to social web mining in the literature [15], [7], [6]. McCallum et al.[6] propose for Author Topic model (AT), as three Bayesian hierarchical models to deal with roles with email datasets. The Author Recipient Topic model (ART) is a directed graphical model which models social role as an explicit graphical model through a latent random variable. A role is therefore a topic mixture characterizing the relation of two persons (that is, the author and the recipient). This work was applied to academic and email networks [6],is improved later on by Rosen-Zvi et al, [7]. Daud et al. [16], focusing on the task of Conference Mining propose for an original model for discovering the latent topics between the authors, venues (conferences or journals) and time simultaneously.

B. Topical Models of Trust

As trust has been the sole focus of artificial intelligence domain, multi- or even inter-disciplinary models of trust are very recent [8]. Models of opinionated trust has been put forth in two distinct but correlated models. Closest to this idea is topical trust [17], [18], using topic labels as edge labels on a social network exemplifying context or nature of a trustworthy relation. More recently, using natural language processing techniques have been leveraged to summarize, integrate or recommend opinion summaries in form of trustworthy topic sets, which would be a part of this contribution. Golbeck and Hendler first set forth the concept of topical trust on the web [17], for applications in trust network building and inference [19] and social recommender systems [20].

While computer scientists can understand topical models, social scientists need better tools to help make sense of that data. Despite popularity of topic modeling to similar problems, word counts and tag clouds are still adopted to interpret information from textual data [21]. For social scientist to be able to leverage topical models for network mining, new models and new metrics need to be proposed [22]. Liu and Fang [22] propose for a tag recommendation algorithm that takes into account users social relations. they model user-created annotations, and the social relations between them using a topic model. They associate each node in this graph with a tag preference vector. They use cosine similarity to find trust and establish links between users and resources through the tags. While Liu [22] et al, focus on combining LDA and trust modeling, distances used are vector-driven, e.g. Cosine similarity, that does not capture latent similarity between two mixtures. While weighing on cold-start problem, Wang et al [23], propose for exploiting tag data to deal with the sparsity problem. Thus to improve the recommendation quality they propose to combine a tag-based neighborhood method with a traditional rating-based CF. To improve the quality of recommendation, they use a tag matrix, which is ultimately used as input to an LDA to generate probabilistic estimation which is jointly given as input to prediction module later on. Their results show improved recommendation quality in turn.

Weng et al. [24], propose for a heuristic measuring the influence of individual twitter users taking both the topical similarity between users and the link structure into account. They utilize an LDA algorithm to distillate and acquire topic sets from twitter users. This is followed by constructing links between twitter users. They show that through existing homophily in twitter, a notion of reciprocity can be observed. While we use the same two steps, A problem with this work is that no information on pre- or post-processing has not been given. Caverlee et al [8], have proposed for Social-Trust++ within which they develop and analyze algorithms for and leveraging community-based notion of trust. While in their modeling they weigh a lot on community model of trust, in order to model and mine implicit communities they emphasize on usefulness of probabilistic topic modeling techniques specifically on LDA. In addition they report that by leveraging LDA-based retrieval, community oriented ranking model results in a significant improvement over other alternatives [25]. As we are using LDA for modeling and inferring relations between users, this work becomes quite close to our idea. Though the Dirichlet distributions are used to model communities rather than individual opinions, as well as the relationship weights are taken as mix of communities, individuals and resources. This is while we model individuals and groups from a trending central topic, and use the distance between their respective latent models as the relationship weights.

III. Mining Topic Facts and Opinions from Social Media

Modern social web is the prominent location for individual and group opinions to be documented and shared. This has attracted marketing businesses to channel existing social media as their marketing playground and promote their ideas or products and in return get feedback from the masses. 70 percent of bloggers are organically talking about brands on their blog, while 38 percent of bloggers post brand or product reviews [26]. Although some social media channels are best for sharing factual messages, others are well utilized for sharing an opinion instead. Since our respective experiment is focused on data gathered from Twitter [1]., from this point forward we focus on content from this social service. While Twitter is an ideal channel for marketing facts, many individuals including bloggers, politicians and celebrities leverage it for sharing their opinions with public. Irregardless of the content being shared on Twitter, timely dissemination of facts and opinions is crucial as it allows building reader confidence through creating a trusted source. Twitters hash-tag feature, similar to tag in tagging services like Delicious [2], allows for a person to define the audience of a message (e.g. $\#bigdata$ will document a fact or opinion to Twitter users whose interest are on large scale analytics).

A. Framework for Topic Mining and Analysis

While the focus of this paper is on learning opinion networks from any gathering of users, its important to realize the obstacles faced when it comes to mining topic sets and their respective distributions from update streams on social services. Main concerns for modeling of topics on social web are twofold: first, the size of the text is often very short, e.g. in case of Twitter only 140 characters, while in the case of Facebook, accessing updates are bound to privacy access rights and often impossible. Second, features are not as focused as a column written in a review website, so saliency and relevancy are always question. For instance a twitter user tweeting about Economics might be talking about his negative experience at a university course while we might be searching for updates related to Eurozone crisis instead. This needs also to be stated that a popular social network is opinionated multi-lingually.

To overcome aforementioned problems, we limit the scope of this manuscript to proposing for using topic models on a tweets surrounding a trending event or product, such as $\#occupy$ movement or $\#iPhone$ product. Topic modeling allows for correlations between topics be found, in addition to the word correlations which constitute topics in the corpora of tweets at hand. This allows for relevant abstract topics to be extracted and pointed out, which ultimately addresses the saliency and relevancy problem, the generative nature of topic models, allows for topics to be inferred from the existing corpora of tweets, which is useful for summarization of a large number of tweet contents. In addition topic modeling can be completely unsupervised, which can easily allow applications to be defined to consume, analyze and summarize streams of updates in real-time fashion. To limit the focus of topics and also overcome the short nature of updates and posts on Twitter for instance, we have chosen also trending issues to make sure that not only we retrieve posts related to our task at hand, but also enough data can be retrieved that does not undermine the performance of algorithm at hand.

Following the justifications of our approach, we have proposed for the following framework 1. First we need to model respective topics surrounding the discussion at hand by acquiring the corresponding features or topics from the corpora of tweets at hand. This is done through a probabilistic topic model. Since we model network of users, we need to do separate modeling of both users and trending topics at hand. While the resulting model helps us to identify common features between the topics in the discussion, it also helps us to eliminate irrelevant topics and allows us to create a main model to compare individual as well as group of users' contribution. To measure such distance, we utilize probabilistic information gains (relative and total) to both analyze the divergent opinions of users towards each other. This will in turn allow us to build resulting opinion matrices (final step of our approach), which in turn can show us divergence of user groups from trending model, or visualize the distance of user opinions from each other in case of user to user comparison. Later on, in experiment part we show case evaluation of a subset of tweets from 2011 where we analyze the results of both matrices using different configurations. In following sections, each part of our framework will be detailed out.

IV. Modeling Tweets

To model the tweets we will use probabilistic topic models [27]. In the following section we detail out what is LDA and how we use it in our task. As pointed out earlier we need to model both the trend (corpora of all tweets) to get an overview of overall opinions, and each and individual user (corresponding users which their tweets are subset of collective corpora at, along with their relevant tweets).

A. Preprocessing

Since we are dealing with user-generated content, prior to any steps to be taken we need to make sure that no user-asserted spelling errors or bad language can get into our way and affect the performance of our algorithm. More over, when dealing with multilingual data it's important to be able to find and filter out tokens of languages under study. Since dealing with lemmatization, segmentation and part-of-speech tagging has been an important problem when analyzing corpora from Twitter, there are existing efforts on this subject [12], [13], [3], [28], [29]. We have used Carnegie Mellon university's TweetNLP [3] tool set [28]. In this tool set authors propose for a tag set, annotated data and features. We used TweetNLP for tokenization and part-of-speech tagging. Figure 2 shows a

[1]Twitter, http://twitter.com

[2]Delicious, http://delicious.com/

[3]TweetNLP,http://www.ark.cs.cmu.edu/TweetNLP/

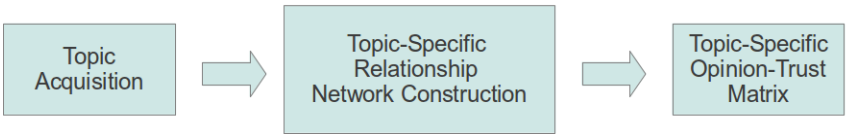

Fig. 1. Overall Framework for Opinion Trust Modeling and Mining: Topic acquisition, involving text preprocessing to remove language faults as well clustering tagged corpora through LDA, followed by dividing resulting distributions into user and trend models. Finally, trust estimation allows resulting topic models to be mapped onto corresponding cells of trust matrix through divergence metrics.

"...Eurozone PMI Services rises to 55.9 in January from 54.2 in December http://bit.ly/xxxx " ..."europe IMF Supports Extension of Greek Loan Repayment Period: Eurozone leaders reportedly reach consensus http://bit.ly/xxxx"

Eurozone ^, PMI ^, Services N, rises V to P 55.9 $ in P January ^ from P 54.2 $ in P December ^ http://bit.ly/xxxx,U ... europe ^ IMF^ Supports ^ ExtensionN of P Greek A Loan N Repayment N Period N : , Eurozone ^ leaders N reportedly R reach V consensus N http://bit.ly/xxxx U

Fig. 2. Preprocessed Tweet lines using TweetNLP: Entry on Top; Tokenized and POS-tagged on Bottom. Tags are presented in distinct colors. For instance,$ tag represents a numerical value, U tag represent a link or URL.

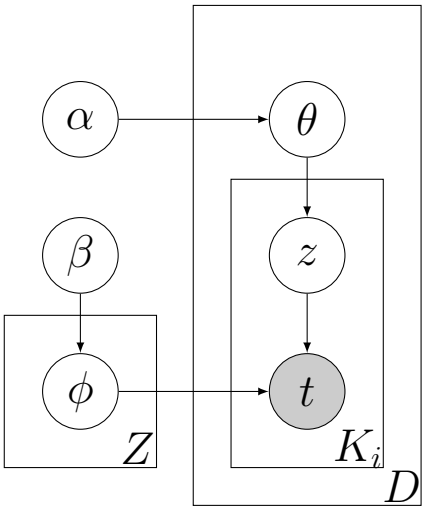

Fig. 3. Graphical Presentation of Latent Dirichlet Allocation

sample tweet (above box) and its resulting processed output (below box).

TweetNLP's Tagger is a Conditional Random Field (CRF) classifier [30], which incorporates arbitrary local features in a log-linear model. The base features includes a feature for each word type, a set of features that checks whether the word contains digits or hyphens, suffix features up to length 3, and features looking at capitalization patterns in the word. This is followed by added features that help leveraging domain-specific properties, unlabeled in-domain data, and external linguistic resources including twitter orthography, traditional tag dictionary and distributional similarity [28]. When we tested our experiment corpus containing 1600 tweets with TweetNLP, base classifier gave a total accuracy of 0.95 (95% confidence interval 0.945 +/- 0.008) which is very decent for such randomly selected corpora. Since the tool set is based on an English dictionary we filtered out non-English tokens easily on second iteration. Although unsupervised POS tagging has been suggested widely in literature for Twitter, our experiment shows that semi-supervised POS-tagging with human annotation could result in more focused results.

B. Acquiring Topics

The goal of the topic acquisition step is automatic identification and extraction of topics that social users are interested in based on the text updates they post. Latent Dirichlet Allocation (LDA) model [4] is an unsupervised machine learning technique that helps identifying latent topic information from large document collection. LDA utilizes bag of words modeling, for categorizing each document with respect to count of vector of words. Based on this assumption, each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over a number of words. An atomic tweet might contain only a single aspect or feature of an event or product, e.g. as opposed to a review of a product. That is why it's important to model the collective opinion of a crowd sharing opinion about a product or an event. These features or aspects could also vary, that is why we have adopted a statistical topic modeling approach to find features or aspects. We used Mallet [4], for modeling topic from existing corpora. which makes use of Gibbs sampling for computing the latent topics. Latent Dirichlet Allocation, models each group of tweets as a mixture of latent topics. Figure 3 shows the graphical notation for LDA. By default of the library, we have used multi-grams [31].

As mentioned earlier, we model two separate groups respectively (resulting in two separate matrices): a generative model for trend, and a generative model for user profiles. Instead of altering the generative process of LDA, we model both models through the same approach but with separate distributions as described as follows.

We take a corpora of K documents, each representing i

[4]Mallet.http://mallet.cs.umass.edu [31]

TABLE I
SAMPLE TOP 5 WORDS IN TOPICS WITH PROPORTIONS FOR TWEETS PRESENTING EUROZONE TREND

Terms	Prob.
economics (6) political (3) reading (2) economic (2) current (2)	0.053
economics (6) make (2) talking (2) politics (2) builders (1)	0.042
economics (9) politics (4) hypnosis (2) year (2) caexpo (2)	0.032
economics (6) pricing (4) effect (3) network (3)	0.023

documents (e..g tweets), such that K_i will be count of all words in corpora in total of d documents.

1) Select $\theta_i \sim Dirichlet(\alpha)$ where $i \in 1, \ldots, D$
2) Select $\phi_t \sim Dirichlet(\beta)$ where $i \in 1, \ldots, K$
3) For each words w_{ij} where $j \in 1, \ldots, N_i$
   - Choose a topic $t_{ij} \sim Multinomnial(\theta_i)$
   - Choose a word $w_{ij} \sim Multinomnial(\phi_{t_{ij}})$

This generative model eventually codes the corresponding trend and respective profiled interests, from which we can infer the unobserved topic and user interest topics through learning model parameters.

Learning process, as mentioned earlier is done through sampling. Learning respective distributions, e.g. the set of topics, their associated word probabilities, the topic of each word, and the particular topic combination of each document) is a problem of Bayesian inference [4]. There are various methods for evaluating LDA and respectively estimating the inferred sets [32]. One of the approaches that currently MALLET[31] is making use of is importance sampling, the generative process for Empirical likelihood evaluation method [6] was made use of in this work. LDA finds a pre-specified set of $|Z|$ topics within $|D|$ documents. Each term t in a tweet with K_i terms then ends up correlated with a topic z. $Z = \{z_1, z_2, z_3, \ldots, z_n\}$ is the set of n latent topics which exemplifies coarseness and resulting final set of topics.

LDA determines a combination of topic sets for each document in the input data throughout the modeling process outlined earlier. Thus, through importance sampling $P(w|\theta^{(d)})$ will be generated as follows:

$$P(w|\theta^{(d)}, \Phi) = \prod\nolimits_n \sum\nolimits_{z_n} P(w_n, z_n|\theta^{(s)}, \Phi) \quad (1)$$

where $P(w|\theta^{(d)})$ is the probability of all multi-grams for a given input document d, z_i is ith latent topic, w_i is ith word of document input D. θ is the document-specific topic distribution probability. $P(w_n|\theta^{(s)})$ are estimated from a synthetic document, randomly-generated using $\theta^{(s)}$. This can be simplified through prior knowledge of d documents before hand thus 1 will be simplified as follows:

$$P(w_i \mid d) = \prod\nolimits_n \sum_{j=1}^{|Z|} P(w_i, z_j \mid d), \quad (2)$$

Following this we can represent latent topics as a list of multi-grams with a probability for each multi-gram indicating the membership degree within the topic. Furthermore, for each document in our corpus we can determine to which topics it belongs, also associated with a degree of membership (topic probability $P(w_i, z_j \mid d)$). An example of two extracted latent topics represented by the top 30 terms is shown in Table I. Beside the terms also the probability for the terms belonging to the topic are shown. For this example we used $|Z| = 30$ latent topics.

C. Divergence Metrics for Trust Modeling

Following the modeling of trend and profiles, two opinion matrices will be made: A matrix for trend-topic divergence (TM), and a matrix for user-user divergence (UM). While the former matrix allows us to be able to measure the distances between individual contributing opinions and the collective opinion from trend model, the latter matrix helps us to be able to measure distances between individual opinions involved in the stream of trending discussions at hand.

As emphasized earlier cosine-based metrics, i.e. $tf - idf$ distance, capture vector distances and no latent information such as saliency, relevancy and polarity which are of high importance in modern information retrieval [33] can be measured in return. Taking this into account for the task of social network analysis through topic modeling, more justified metrics are needed. In probabilistic information theory, Kullback-Leibler Divergence, or simply relative divergence, is of high importance when evaluating probabilistic information retrieval tasks, as it can measure the distance between two resulting distributions from topic models. Since measuring distances of collective opinion of groups, or individuals on a trending ground, can be modeled through information divergence. Kullback-Leibler divergence estimates the number of additional bits needed to encode the distribution Q, using an optimal code for P, and having a combined vocabulary size of $|Z'|$; in our case the number of latent topics $|Z|$ times the number of rating classes $|R|$.

$$D_{KL}(Q||P) = H(Q;P) - H(Q) = \sum_{i=1}^{|Z'|} q_i * Log_2(\frac{q_i}{p_i}) \quad (3)$$

A problem with adopting divergence metrics is non-symmetric nature of this metric, e.g. $D_{KL}(Q||P) \neq D_{KL}(P||Q)$. This nature of divergence metrics can be used to model directed social networks. Then $D_{KL}(Q||P)$ can model a network of two nodes Q and P, with a directed edge from Q towards P weighting as $|D_{KL}(Q||P)|$. In addition to lack of symmetry, a latter objection to using Kullback-Leibler as a metric is lack of normalization, as the resulting values might fall between any range of numbers.

This is while a weighted graph needs a normalized weight to present the distances between any set of nodes on the resulting network. Thus instead of relative divergence, total divergence of two respective distributions can be used instead. Thus we use Jensen-Shannon Divergence to model similarity between two respective distributions at hand, as a result with two distributions of Q and P:

$$D_{JS}(Q||P) = \frac{1}{2}(D_{KL}(Q\|M) + D_{KL}(P\|M)) \quad (4)$$

M is the average of two probability distributions and is calculated as $M = \frac{1}{2}(Q + P)$ and $D_{KL}(Q||P)$ is Kullback-Leibler divergence of Q from P calculated using Eq.3. We further normalize the result of Eq.4. We can define topical distance of two profiled users U and V, as *divergent opinion trust* between U and V :

$$trust_{u,v} = 2 * \sqrt{D_{JS}(U,V)^2} \quad (5)$$

where $trust_{u,v}$ is trust between users U and V, which is in turn measured through normalized distance of D_{JS}. Using Jensen-Shannon Divergence as a metric gives two benefits over Kullback-Leibler: since the results are normalized they can be mapped onto continuous range of [0,1], moreover since Jensen-Shannon is symmetric we can model free-form, .e.g undirected, edges on resulting networks that we model through our framework as result.

$$trust_{u,v} = \begin{cases} 0, & D_{JS}(U||V) = 0 \\ 2 * \sqrt{D_{JS}(U,V)^2}, & D_{JS}(U||V) \geq 0 \end{cases} \quad (6)$$

Divergence metrics either find a latent-level similarity between two distributions or not, thus allowing the trust levels over these relations be uniformly distributed.

V. Experiment: Mining Networks of Eurozone Trending News Corpora

As a proof of concept, in this section we are presenting our experiment with a Twitter corpus. To begin with, first we present our dataset briefly, followed by evaluation section where we detail out the evaluation results with respect to corpus and data at hand. Our analysis will be focused on the features of generated networks from perspectives of social network analysis.

A. Dataset

We have used Tweets 2011[34] dataset, which is part of TREC 2011 Microblogs Track. This dataset contains identifiers to more than 16 million tweets and serves as a realistic representation of Twitterosphere as dataset is not reprocessed, nor it has been normalized to filter out Spam for instance. Since data is gathered in 2011, we decided to focus on trending subjects from 2011 year with prior knowledge of trending events or subjects on Twitter. Among subjects we decided to focus on a trending news story, namely Eurozone's economic crisis. Other subjects are under study and are subject to publication for future work. To make sure that we pull out corresponding tweets we expanded the queries with hash-tags related surrounding $Eurozone$. We have used an on-line hash-tag search engine, i.e. Hashonomy [5], that analyzes trends from Twitter to gather related hash-tags to our work. As a software stack is being developed around our current framework we plan to use web semantics for query expansion by analyzing existing twitter content, at hand. After query expansion, the retrieved results grossed to 1600 distinct tweets. After network extraction a total of 695 user profiles were extracted. We additionally expanded the profiles with related tweets made by the same person to increase the performance of algorithm as well.

B. Evaluation

User matrix which is the latter product, and perhaps most important output,of the framework presented to you so far is a two-dimensional matrix storing $< user, user >$ pairs with their divergences associated. Dimensions are based on the size of the users extracted from the corpora under focus. Similar to previous part, we squeeze the matrix to various sizes based on the size of networks we are interested in extracting. We follow the same methodology by grouping users into fixed sizes of clusters. Now being capable of generating generating networks of various diameters, we use two social network analysis metrics to study the resulting networks, on local and global scales [35]. Fig 4 showcases visualizations of four different diameters of resulting networks using Gephi[6].

With respect to node level analytics, average weighted degree of a node is studied. we know that count of edges attached to nodes is an effective measure of importance of nodes. The higher the value, the more important a node is in a graph. Proportion of nodes directly connected in the entire graph is as a result measures the reachability of nodes. The weighted degree of node i is simply the total of values w_{ij} associated with L links in total, as follows:

$$k_i = C_D(i) = \sum_{j}^{L} w_{ij} \quad (7)$$

With respect to network level analysis, we have used Clustering coefficient which is a measure of degree to which nodes intend to cluster together. The clustering coefficient for the whole network is given by Watts and Strogatz [36] as average of the local clustering coefficients of all the links n as follows: groups characterised by a relatively high density of ties

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i \quad (8)$$

Where C_i is local clustering coefficient calculated as follows:

$$C_i = \frac{\{e_{jk}\}}{k_i(k_i - 1)} \quad (9)$$

Following equations 7, 8, 9, results of social network metrics for graphs generated from user matrix (UM) are plotted in Fig.5 and Fig.6.

[5]Hashonomy,www.hashonomy.com

[6]Gephi,www.gephi.org.

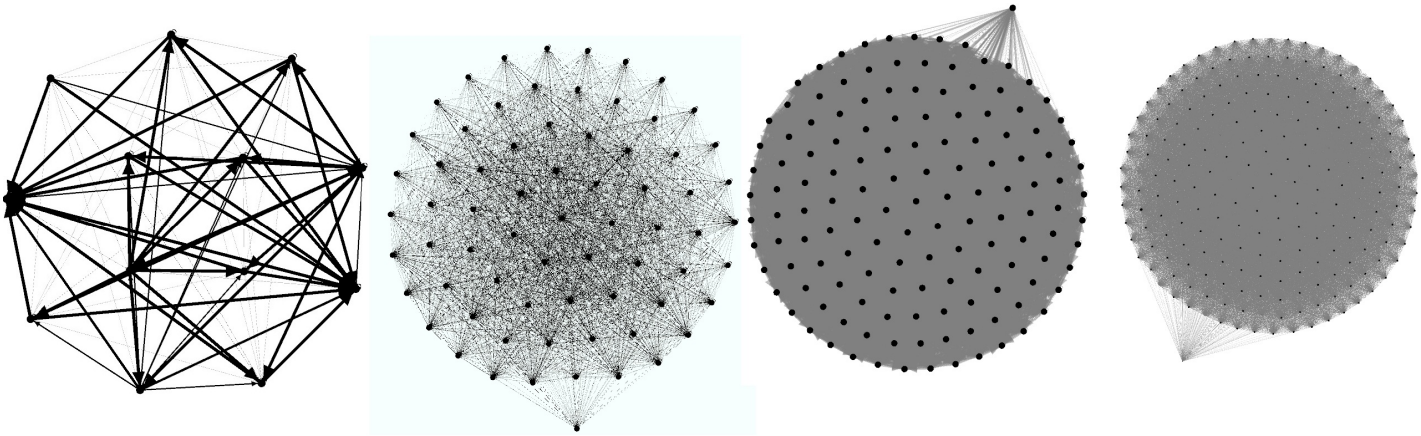

Fig. 4. Fruchterman-Reingold visualization of evolution of learned trust network of Eurozone twitter(er)s: from left to right, network of user cluster of 2%,10%,20% and 30% size. Weights on the edges of graph are rescaled to reflect the impact of divergence. Larger sizes of networks were not presented as they lose their structural visibility. Weight values are presented using visualized pressures on network links.

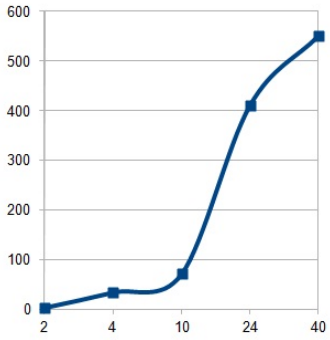

Fig. 5. Node level analytics on trust graph: average sum of weights of the edges (average Weighted degree). Horizontal axis presents the percentage of total user groups sampled for the experiment, while vertical axis plots the respective degree.

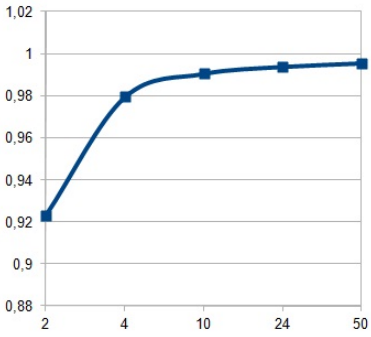

Fig. 6. Network level analytics on trust graph: degree of clustering of nodes (Clustering coefficient). Horizontal axis presents the percentage of total user groups sampled for the experiment, while vertical axis plots the respective coefficient.

It is generally accepted that if a social graph is to reflect characteristics of real-world networks, within its structure nodes will tend to create tightly knit groups characterized by a relatively high density of ties.

Taking into consideration this hypothesis and focusing on the plotted results for various densities of user clusters, we can observe easily that weights on the first plot, that the average weighted degree sharply increases for less than 50% of networks. This claim can easier be justified through clustering coefficients, as even a small diameter network is densely clustered (starting from 0.923). As a result, we can easily observe that networks generated via our framework are much suitable to model real-world social networks. Moreover probabilistic generative process of our algorithm, never leaves a distance empty unless two profiles are completely distant from each other.

VI. Conclusion and Future Work

Within this paper we have proposed a framework for opinion-mining from Twitter's content corpora, through which latent topics are acquired and then used for generating opinion trust matrices. These matrices are then used to generate weighted social networks. An analysis presented showed that these networks represent real-world models of profiles and trending news or events that can be used for applications of business intelligence such as advise giving, viral analytics and influence metrics for instance. Being an initial analysis, we are planning to study more trending stories and events to better establish the concept set forth in this manuscript. Since it was evident that data size can affect the performance of algorithm, we are planning to use larger data sizes as a future work. This is followed by leveraging resulting networks for ranking, recommendation and summarization tasks.

References

[1] D. Boyd and N. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of computer mediated communication-electronic edition*, vol. 13, no. 1, p. 210, 2007.

[2] N. P. Hummon and P. Doreian, *Computational Social Network Analysis*. Springer London, 2010, vol. 12, no. 4, ch. 11, pp. 2–25.

[3] A. Pak, "Twitter as a corpus for sentiment analysis and opinion mining," *Proceedings of LREC*, pp. 1320–1326, 2010.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, January 2003.

[5] R. Krestel and N. Dokoohaki, "Diversifying product review rankings: Getting the full picture," in *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. Washington, DC, USA: IEEE, Aug 2011, pp. 138–145.

[6] A. McCallum, X. Wang, and A. Corrada-Emmanuel, "Topic and role discovery in social networks with experiments on enron and academic email," *Journal of Artificial Intelligence Research*, vol. 30, no. 1, pp. 249–272, 2007.

[7] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers, "Learning author-topic models from text corpora," *ACM Transactions on Information Systems*, vol. 28, no. 1, pp. 1–38, 2010.

[8] J. Caverlee, L. Liu, and S. Webb, "The socialtrust framework for trusted social information management: Architecture and algorithms," *Information Sciences*, vol. 180, no. 1, pp. 95–112, 2010.

[9] H. Kim, K. Ganesan, P. Sondhi, and C. Zhai, "Comprehensive review of opinion summarization," Illinois Environment for Access to Learning and Scholarship, Tech. Rep., 2011.

[10] H. Chen and D. Zimbra, *AI and Opinion Mining*, May 2010, vol. 25, no. 3, pp. 74–80.

[11] X. Zhou, Y. Xu, Y. Li, A. Josang, and C. Cox, "The state-of-the-art in personalized recommender systems for social networking," *Artificial Intelligence Review*, vol. 37, no. 2, pp. 119–132, May 2011.

[12] A. Bifet, "Sentiment knowledge discovery in twitter streaming data," *Discovery Science*, pp. 1–15, 2010.

[13] B. OConnor and M. Krieger, "Tweetmotif: Exploratory search and topic summarization for twitter," *Proceedings of ICWSM*, pp. 384–385, 2010.

[14] T. Hofmann, "Probabilistic latent semantic indexing," *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 99*, vol. pages, pp. 50–57, 1999.

[15] M. Forestier, A. Stavrianou, J. Velcin, and D. Zighed, "Roles in social networks: Methodologies and research issues," *Web Intelligence and Agent Systems*, vol. 10, no. 1, p. 117133, 2012.

[16] A. Daud, J. Li, L. Zhou, and F. Muhammad, "Conference mining via generalized topic modeling," *Machine Learning and Knowledge Discovery in Databases*, vol. 5781, pp. 244–259, 2009.

[17] J. Golbeck and J. Hendler, *Accuracy of Metrics for Inferring Trust and Reputation in Semantic Web-Based Social Networks*, 2004, pp. 116–131.

[18] T. Knap and I. Mlynkov, "Towards topic-based trust in social networks," *Ubiquitous Intelligence and Computing*, p. 635649, 2010.

[19] D. Godoy and A. Amandi, "Enabling topic-level trust for collaborative information sharing," *Personal and Ubiquitous Computing*, Aug 2011.

[20] A. Zarghami, S. Fazeli, N. Dokoohaki, and M. Matskin, "Social trust-aware recommendation system: A t-index approach," in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 3. IEEE Computer Society, 2009, pp. 85–90.

[21] D. Ramage, E. Rosen, J. Chuang, C. Manning, and D. McFarland, *Topic modeling for the social sciences*, 2009, pp. 2–5.

[22] K. Liu and B. Fang, "Integrating social relations into personalized tag recommendation," *2010 Second International Conference on Intelligent Human-Machine Systems and Cybernetics*, pp. 292–295, Aug 2010.

[23] Z. Wang, Y. Wang, and H. Wu, *Tags Meet Ratings : Improving Collaborative Filtering with Tag-Based Neighborhood Method*. ACM Press, 2010.

[24] J. Weng, E. Lim, J. Jiang, and Q. He, *Twitterrank: finding topic-sensitive influential twitterers*. ACM, 2010, p. 261270.

[25] S. Kashoob, J. Caverlee, and K. Kamath, *Community-based ranking of the social web*. ACM Press, 2010, p. 141.

[26] Technorati, "Technorati state of blogosphere 2010," Tech. Rep., 2010. [Online]. Available: http://technorati.com/state-of-the-blogosphere

[27] R. Krestel and P. Fankhauser, "Tag recommendation using probabilistic topic models," *ECML PKDD Discovery Challenge 2009 (DC09)*, p. 131.

[28] K. Gimpel, N. Schneider, B. O. Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: Annotation , features , and experiments," *Most*, no. 2, pp. 42–47, 2011.

[29] L. Cagliero and A. Fiori, *Analyzing Twitter User Behaviors and Topic Trends by Exploiting Dynamic Rules*, 3rd ed. Springer London, 2012, pp. 267–287.

[30] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Rapport technique MSCIS0421 Department of Computer and Information Science University of Pennsylvania*, vol. 50, no. 7, p. 90, 2010.

[31] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002. [Online]. Available: http://mallet.cs.umass.edu

[32] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," *Proceedings of the 26th Annual International Conference on Machine Learning ICML 09*, vol. 382, no. d, pp. 1–8, 2009.

[33] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '08. New York, NY, USA: ACM, 2008, pp. 659–666.

[34] NIST, "Tweets2011: Trec 2011 microblog dataset," 2011, http://trec.nist.gov/data/tweets/. [Online]. Available: http://trec.nist.gov/data/tweets/

[35] A. Marin and B. Wellman, *Handbook of Social Network Analysis*, P. Carrington and J. Scott, Eds. Sage, 2010.

[36] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *The Structure and Dynamics of Networks*, vol. 393, no. 6684, p. 440442, 2006.