Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 50 (3D) 119-125 (2018)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

A Cross-Domain Analysis for Actor Identification in Multiple Networks

F. NOOR[++], A. SHAH, S. A. KHAN

Yanbu University College (YUC), Saudia Arabia

Received 10th June 2018 and Revised 15th September 2018

Abstract: This paper proposes a technique for actor identification in any individual's multiple communications network. This problem is specially an important problem for personal overview, intelligence gathering and countering terrorism but can also be applied on automatic actor identification for social network managements of any individual. The technique performs the correlation of different features using the feature set consisting of basic features, text segment feature and authorship attribution features. The features are extracted from datasets of three different ego-centric networks (Facebook, Messages and Email), collected over a period of three months. The preliminary result shows effectiveness of the technique in identification of actors in a communication network

Keywords: Actor identification, Communication Networks, Ego-Centric Network Analysis, Cross correlation

1. INTRODUCTION

Online communication media is a means of interactions among people in which they create, share, trade and assess data and thoughts in groups, communities and networks. Whenever people interact directly or indirectly with other people or communities using online media, interaction networks are created. These interaction networks provide lots of information about communication patterns, location information, friendships, and other interaction arrangements. In the current era, people are becoming dependent on online network sites and massive online data is generated on daily basis. It is very complex to analyse this huge data for specific purpose manually, thus results in the emergence of computational means of analysis.

To analyse interaction networks different network architectures are being used. In common practice, networks are analysed standalone as monoplex networks. However, multiple interaction connections can also be analysed using advanced multi-layer model. In multi-layer paradigm, each communication network is considered as one dimension and is modelled in a single layer. Henceforth, the multi-layermodel of multiple different connections has L layers in addition to nodes v represented in set V and edges e in set E. The node vrepresents the different online accounts and edge e represents the communication between those nodes.

In identification of actor, studies have exploited all possible node based attributes and edge based attributes such as profile attributes, content or frequency of communication, and communication based attributes to dig out the best combination. In dark networks studies, use of simple actor features like name, location, preferences cannot provide enough satisfactory results hence analysis of link (i.e. communication) is an important task to provide the main features. However in online networks normally the individual are not interested in hiding their information, their roles and their importance but sometime uses different nickname and random information for the sake of signup or their context changes with time or they changes info according to need of network. In all such cases, text analysis of their communication can provide the additional features for identification of individuals on different networks especially those where basic profile info is limited and media are text based.

This paper proposes a technique of identifying actors across different online communication networks using multi-layer paradigm. The identification will be based on the correlation of communication patterns and analysis of the contents in communication using natural language processing tools along with other basic features. The problem is important for establishing a 1-1 correspondence of actors across multiple communication network domains in any individual personal multiple networks.

2. LITERATURE REVIEW

Network Analysis is a compilation of methods used to identify and analyse patterns in network systems. It is also been exploited for illegal activities such as for planning and launching terrorism (Fu *et al.,* 2012), criminal activities coordination & organization (Whissell, 1989). Now days, law enforcement agencies are using online communication media as an

[++]Corresponding author :fozia noor khan@yahoo.com.

informative source for investigation (Alderson, 2011), (Marshall *et al.,* 2004). They are analysing communication media content for the purpose of illegal data or activity detection and tracking. (Fu *et al.,* 2012) proposed six-element analysis method for terrorist activities based on online community network. (Erlin *et al.,* 2008), proposed a concept to integrate between content analysis (CA) and network analysis (SNA). In their approach, they proposed a method to analyse communication transcripts to filter out related messages from unrelated messages.

In online communication networks, the extensive research is done in previous few years to explore the possibility of performing entity resolution (identity resolution) on online communication networks. Researchers such as (Labitzke *et al.,* 2011) studied VZ, Facebook, Myspace, Xing to resolve identities. Narayanan (Narayanan *et al.,* 2009) explored Twitter to match friends. Different researchers (Irani *et al.,* 2009), (Malhotra *et al.,* 2012), (Perito *et al.,* 2011) studied diverse networks i.e. Facebook, Myspace, Twitter, BlogSpot, LinkedIn and explored basic profile attributes like full name, school, birth year, city, location and age to resolve identities. However, the approach of using basic attributes has its limitation, as it is not enforced that actors will use same particulars on all networks. Some researchers also exploited the online connections and content based attributes in studies along with basic attributes or as standalone attributes in matching actors.

Automated text analysis and sentiment analysis are the computational means of analysing & judging the text and are currently playing an important potential role in reputation and government intelligence, review related technologies and e-politics. Along with all these general uses, analysis of online messages and comments can used to identify the messages that have inappropriate or violent information and to identify groups and individuals involved in illegal activities (Fu *et al.,* 2012). Researchers (Goga *et al.,* 2012), (Gani *et al.,* 2012), (Iofciu *et al.*, 2011), (Szomszor *et al.,* 2008) working on identity resolution across multiple communication domains have also exploited content of the messages posted to extract important information like phone number, location, names and other attributes.

The content (posts, movies, wall messages) created by user on online networks and characteristics of content created (author, tag, URL's stylistic features) has also been explored by researchers (Sattikar, 2012), (Jain *et al,* 2011). (Goga *et al.,* 2013) resolved 94.7% of Twitter and Yelp identities by using timestamps of content and author's writing style. In online tagging network (Delicious and Flickr), (Tereza, 2011) uses tags to resolve the identities between networks.

Lot of researches have been done with the focus on the online blogging or networking sites i.e. Facebook, twitter etc. The communication network including Emails and chat Messaging have left alone. This research has a focus on these left out networks along with one of the famous OSN Facebook. In particular, this research will analyse and embed existing natural language processing techniques along with network measures to the individual identification architecture.

3. PROPOSED MODEL

The proposed model has four main modules named as pre-processing module, visualization module, feature preparation module and individual identification module. **(Fig. 1)** illustrates the proposed model diagram. The pre-processing Module is responsible for cleaning incomplete, noisy data. The data is modelled as network and is visualized for any obvious pattern detection. The network measures are also calculated at this stage. Feature preparation module extracts the basic feature and performs text analysis of the communication content to extract different text based features. Until the end of this stage,

$$F = \{ Fb , Ft, Fn \}$$

Where Fb represent basic features, Ft represent text based features and Fn represent network measures. These features are sent to individual identification agent for matching. Individual identification agent uses the syntactic and semantic methods depending on the type of feature to match attributes. This individual identification agent designed as a correlation based template-matching agent.

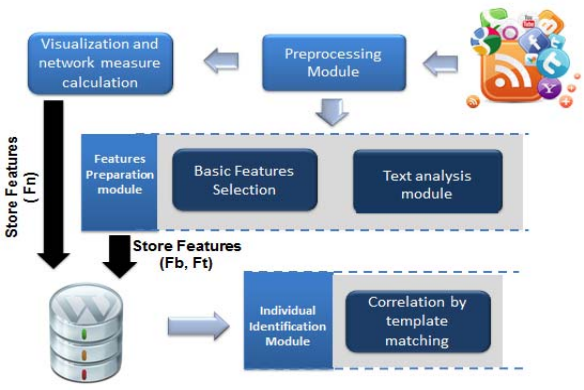

Fig. 1: System Model

3.1. Data Collection

Important problem within network analysis is to collect and extract useful data. The interpretation of the online communication data and identification of important patterns of the network directly depend on the quality of data used in analysis. Data sets are publically available but due to technical or privacy reasons only the existence of communication between nodes are given. The content of those exchanges are not known,

except for Enron and Henry Clinton public email communication dataset. However, this research needs the dataset of same individual for more than one communication network as shown in **(Fig.** 1**)**.

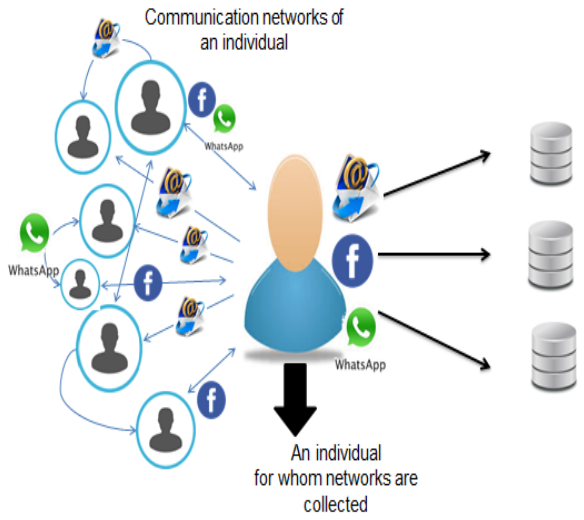

Fig. 2: Data collection scenario

Thus for this study the only way to get a realistic dataset is to use real data. To perform experimentation, communication domains considered are Emails, SMS/WhatsApp and Facebook. An egocentric network is collected over a period of three months from these network.

3.2. Pre-processing Module

After data is gathered, pre-processing module works to prepare the data for analysis. The entities, whose attributes are more than 40% missing, have discarded. All names converted in to lower case. The module also checks the duplicate entries and removes them to avoid redundancy. The unconnected nodes are also removed during cleaning of datasets.

3.3. Modelling the networks

Different ways are used to model the network. Most common representations are adjacency matrix and node-edge diagram. Adjacency matrix representation has been used extensively for network analysis (Ghoniem *et al.,* 2004). They represent the mono-plex layer network mathematically by representing the relationship among nodes. However, multilayer network includes multiple network layers (dimensions). Thus simple adjacency matrix cannot represent all of them simultaneously. Multi-dimensional network not only includes the edges between the nodes within one dimension (intra-edges) but also includes edges between the nodes of different dimension (inter-edges). Henceforth supra-adjacency matrixis used to represent them.

Visualization using node-edge diagram has a rich history and used as an analytical tool (Freeman, 2000). Many stand-alone tools developed to model node-edge diagram cover only mono-plex network paradigm. They represents the multi-dimensions of multi-layer networks in single graph connecting nodes with all possible relations using different colours or separately as independent monoplex networks.

In current approach, we have modelled multiple networks into the multi-layer paradigms and have used muxViz (Domenico, 2015) to generate multi-layer graph to visualize and analyse the datasets. In our case the networks in multi-layer network only have intra-edges but they are analysed in multilayer paradigm to identify existing visual patterns and identifications among the three different networks. The multi-layer visualization is shown in **(Fig.** 1**)**

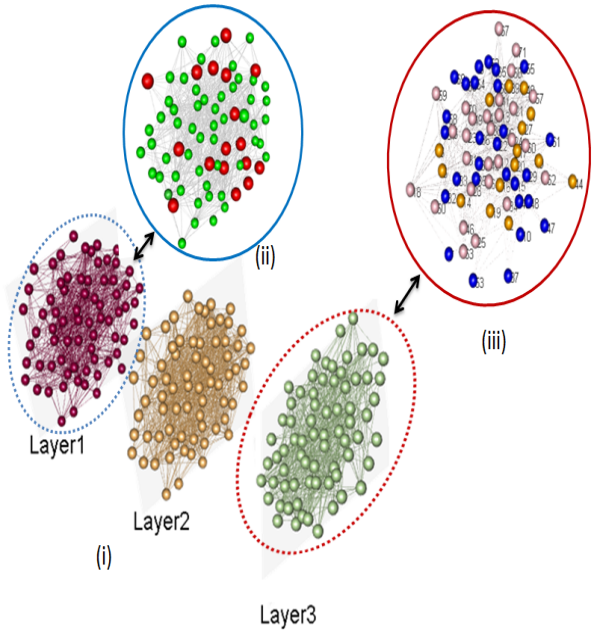

Fig. 3 Multi-layer visualization (i) three dimensions of network (ii) Top nodes represented in red in dimension 1 (ii) Eigen centrality of dimension 3 nodes represented using 3 colour spectrum

Basic network measures like degree centrality, betweenness centrality, eigenvector centrality, closeness centrality and clustering coefficient is calculated to use as feature in the matching. Each centrality measure provides information of actors from different perspectives used to find the role of actors in the networks.

3.4. Feature preparationModule

Feature preparation module is responsible to prepare and select relevant features for the model. This module extracts the basic feature and performs the text analysis to extract text-based measures.

3.4.1. Basic feature selection

The basic features depend on the online communication media selected. The basic features of the communication media like Facebook, Twitter includes profile based attributes such as name, gender, city, date of birth, location etc. In our case two of the networks selected are email and messaging network that do not include this kind of basic information. These networks register on either email addresses and/or phone number. The Facebook includes the complete profile information but to include the common basic features the feature set $Fb(v)$ include email, name, nickname, profile photo and phone number for every node.

3.4.2. Text based features extraction

The content individual shares with (writes to) each other is of great importance and can play an important role in individual identification. (

Fig. 4) shows the feature preparation module, which process the crawled text data to generates tokens of interest i.e. words, phrases, symbols, or other meaningful elements. Three types of features generated by text analysis for every node and is stored in text based feature list $Ft(v) = \{ft_1, ft_2, ft_3\}$.

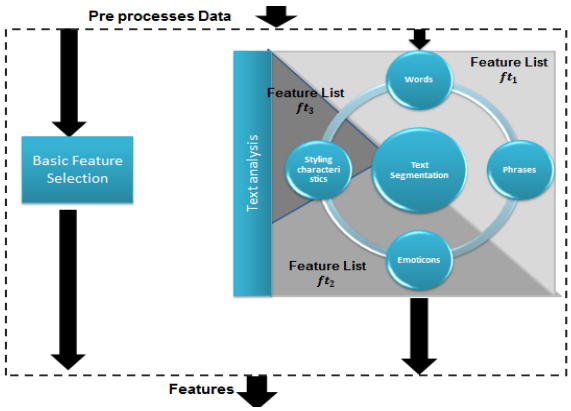

Fig. 4: Text Analysis Module

First list ft_1 contains the meaning-full words and phrases, generated by tokenizing the text using non-letters to convert text in to words. All the tokens are transformed in to lower case letter. In next step, stop words are removed and to kens are stemmed into root words. This research is limited to only English letter fonts to avoid the tokenization issues of different languages. **(**

Fig. 45) shows the text segmentation process used in Rapid Miner. While the **(**

Fig. 46) shows the calculation of different vocabulary size using python for a sample node.

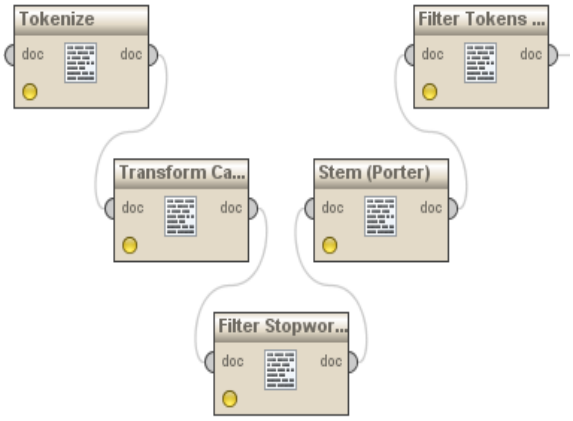

Fig. 5: Text tokenization in Rapid Miner

```
172 99 62 Email_Abdullah/20130918-DINNER TOMORROW-601.txt
246 125 92 Email_Abdullah/20131003-FW_ Accounts details-570.txt
95 70 48 Email_Abdullah/20131006-FW_ Kitty Money-569.txt
47 39 23 Email_Abdullah/20131209-links to grammer check-395.txt
373 210 154 Email_Abdullah/20131221-FW_ Our new member_s office extension-336.txt
42 30 15 Email_Abdullah/20140113-print plz-256.txt
164 106 74 Email_Abdullah/20140114-CDC First Meeting-251.txt
54 36 20 Email_Abdullah/20140115-Fwd_ Mother In Law Passport-243.txt
57 45 34 Email_Abdullah/20140119-Missing You-237.txt
134 97 73 Email_Abdullah/20140128-finalization of MISSION_ Vision of CDC-226.txt
976 407 306 Email_Abdullah/20140129-FW_ TEXTBOOK STATUS AT YUC WAREHOUSE-225.txt
168 97 65 Email_Abdullah/20140202-Telephone Numbers-219.txt
101 75 51 Email_Abdullah/20140205-Your Availability-214.txt
158 113 79 Email_Abdullah/20140206-Teacher _sTextBooks for semester 132-212.txt
82 64 43 Email_Abdullah/20140210-CDC 2nd meeting-198.txt
176 97 62 Email_Abdullah/20140210-Re_invitation for BHAI_s-195.txt
172 114 80 Email_Abdullah/20140212-CDC 2nd meetings-192.txt
```

Fig. 6:Samples of extracted emails and their calculated vocabulary size

Second list ft_2 contains emoticons and chat slangs. This list of vocabulary can help in identifying actors from his style of writing specifically chat messages. Emoticons are not easy to identify. For identifying emoticons and slangs, list of emoticons and slangs from Wikipedia is used. Code to extract the slangs and emoticons has written in python that uses regular expression for finding emoticon and slangs. List ft_2containing emoticon and slangsis generated for each node(v) in the network.

The average word length, average sentence lengths, average unique items used and average term repetition frequencies, percentage of stop word, and percentage of punctuations are computed and stored in list ft_3. This features set reflects the authorship style and can help in identification. The summary of three sublist (ft_1, ft_2, ft_3) of text-based feature $Ft(v)$ is given in **Fig.** 7.

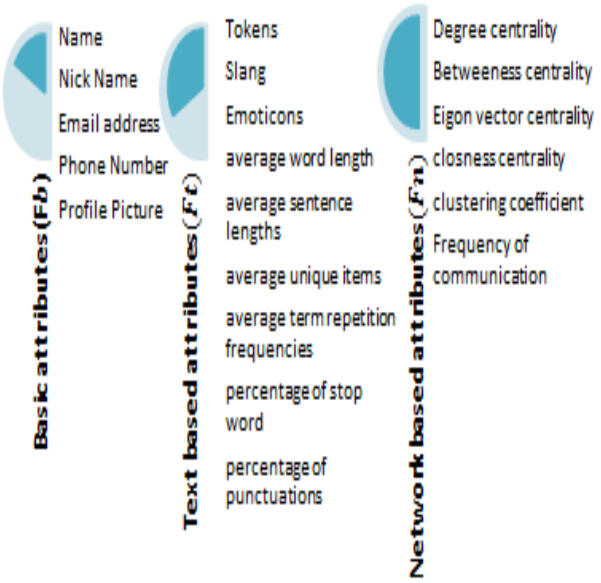

Fig. 7: Three different feature sets *F(l)*

Weights W_l assignment is the last step of finalizing the features using hit and trial method. Weights are assigned to each set of features(i.e. basic, text based and network) carrying the total weight of 1. Attributes with high weight are given more importance in decision making than others. Highest weight age is given to basic attributes then text based attributes then network based attributes.

3.5. Individual Identification Module

Individual identification module based on features set provided performs the cross correlation. Correlation gives the measure of the degree of similarity between actors/nodes as template matching performed on a node-by-node basis on all domains data, under consideration. Nodes having highest degree of similarity identified as same actor. Eq.1 below shows the mathematical equation for the cross correlation process.

$$\max_j \sum_{l=1}^{3} W_l \left(F_{lk}(vi) \quad \oplus F_{lk}(vj) \quad \right) \forall j = 1..n \text{ Eq.(1)}$$

Where $i = \{1\ to\ N\}$ and $j = \{1\ to\ M\}$ are two different network layers having different number of actors in the network, F_l represent the feature subset and k represent the features in the subset while $W_l$is weight faction of feature set F_l.

Different syntactic and semantic techniques are used to calculating similarity measure based on the type of feature correlated. Syntactic technique i.e. Jaro distance (Jaro, 1978) is used to correlate the names of the user. While email, phone are compared using simple string comparison functions with boolean output. Different researchers like Marshall (2004) also perform identity matching by strict comparison between first name, last name and date of birth attributes. It is not a good idea to use strict comparison method where attributes can have minor differences as this method lacks the flexibility thus may lead to false negatives. However, in our case even a minor change in phone number and emails means a different individual. Token-based attributes are used as bag of words and similarity score depend on many items being shared. In network-based feature, the difference in centrality is considered as a distance between two. Difference of Frequency of communication is also considered as distance. Each attribute of actors from one network is correlated with the corresponding attribute of actor in other network. The resulting similarity score is a value between 0 and 1 for each feature.

In the next step all similarity scores inside a single set of features are summed to develop the similarity score of pair of profiles for a single feature set F(l). The similarity score received are not normalized, and every feature set carries its own score. The total similarity score of each feature set depends on the number of features in the set. The normalization of score can give a uniform scale to quantify the score of each feature sets. The raw score and normalized score carry the same effect, but normalized score sum up to the sum of irrespective of total features number, thus limits the score set $(0 \leq \text{lw} \leq 1)$.

The normalized similarity score of each feature set are multiplied with weight assigned wl. Weighted similarity scores of each set of feature are summed up. The similarity scores that are above a threshold value are considered as match. The similarity scores that are below a threshold value are considered as no-match. While the similarity scores equals to threshold is possible-match. The validation of the framework is manually done to ensure correct results, as there is no trustable public dataset to be used as ground truth is available

4. RESULTS

The approach is tested using two cases. In both cases, data is sampled first to simulate this approach. We select 40% nodes from data set. In CASE I, dataset of Facebook is used as labelled data and entities are matched in rest of two dataset using features. To practice CASE 2, dataset of Email is used as labelled data and other two are used as unlabelled datasets.

In CASE I where Facebook dataset is labelled78% of entities can be matched in messages dataset while only 60% entities are matched in email. While in CASE 2if Email dataset is labelled then around 79% of the entities are matched in Messages but 58% entities are matched in the Facebook. The results are shown in **(**

Fig. 48) The results are still preliminary but provide a great opportunity to explore individual identification in depth. It also provides further opportunity to analyse the role of each feature set on the analysis.

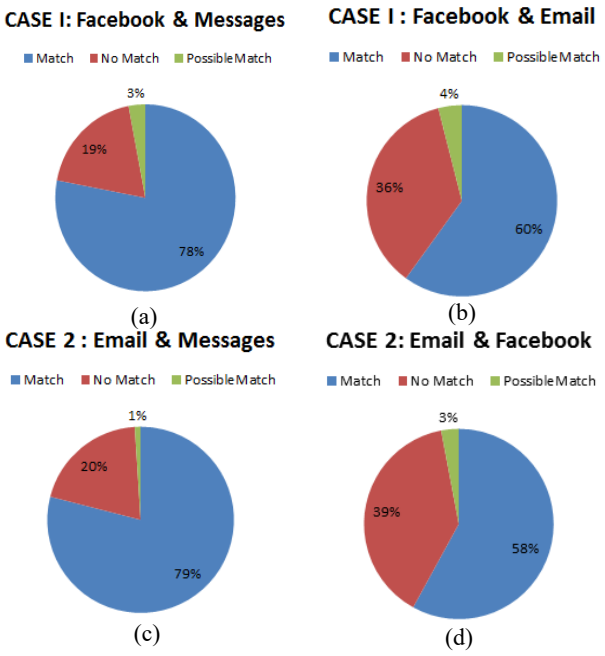

Fig. 8 CASE I and CASE 2 initial results

5. CONCLUSION

This study approach, identify actors across multiple communication domains using different type of feature such as profiles based features, individual's communication link features and node measure. The identification process works on similarity measure, calculated by correlating nodes using list of features. The spotlight of this paper is the use of text analysis in order to extract the features of communication between nodes. The results presented in this paper are preliminary and have a great opportunity of further extension.

REFERENCES:

Alderson. M, (2011) "Facebook: a useful tool for police?" Connected cops. 25 January 2011. Web. 3,

De Domenico, M. , M. A. Porter, A. Arenas. (,2015) Multilayer Analysis and Visualization of Networks, published .

Erlin, Y. N, and A. Rahman. (2008) "Integrating content analysis and SNA for analyzing asynchronous discussion forum," Information Technology, 2008. ITSim International Symposium on 26-28 vol. 3, 1–8,.

Fu, F. J. , J. Chai, and S. Wangl., (2012) "Multi-factor analysis of terrorist activities based on network," Intelligence and Financial Engineering (BIFE), 2012 Fifth International Conference on 18-21 476–480,

Freeman, L. C. (2000). Visualizing S. networks. Journal of S. S., 1–15.

Goga, O., H. Lei, S. H. K. Parthasarathi, G. Friedland, RobinSommer, and R. Teixeira, (2012). On exploiting Innocuous User Activity for Correlating Accounts across Network Sites,"

Gani, K., H. Hacid, and R. Skraba, (2012) Towards multiple identity detection in soc networks," in Proceedings of international conference companion on World Wide Web,.

Ghoniem, M..J. Fekete, (2004) A Comparison of Reada bility of Graphs Using NodeLink and Matrix Represent ations. IEEE Symposium on information visualization Austin, TX,

Iofciu, T., P. Fankhauser, F. Abel, and K. Bischo_, (2011) Identifying Users Across Tagging Systems," in ICWSM,.

Irani, D., S. Webb, K. Li, and C. Pu, (2009) Large Online S. Footprints-An Emerging Threat," in Proceedings of the International Conference on Computational Sci. Engineering – Vol. 03, ser. CSE,.

Iofciu, T. P. Fankhauser, F. Abel, and K. Bischoff. (2012) Identifying Users Across S. Tagging Systems. In Proceedings of the 5th International AAAI Conference on Weblogs and Media, ICWSM '11, 522-525.

Jaro, M. A. (1978) A record linkage system: Users manual. Bureau of the Census,.

Jain, P. , T. Rodrigues, G. Magno, P. Kumaraguru, and V. Almeida. (2011) Cross-Pollination of Information in Online S. Media: A Case Study on Popular S. Networks.In Proceedings of the IEEE, 3rd International Conference on S. Computing,

Labitzke, S. , I. Taranu, and H. Hartenstein, (2011) What your friends tell others about you: Low costlinkability of network profiles," ser. SNAKDD,.

Malhotra, A. , L. Totti, W. Meira, P. Kumaraguru, and V. Almeida, (2012) Studying User Footprints in Different Online S. Networks," International Workshop on Cybersecurity of Online S. Network (CSOSN),

Marshall, B. S. Kaza, J. Xu, T. Petersen, C. Violette, (2004) "Cross-jurisdictional criminal activity networks to support border and transportation security," in Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on, 100-105.

Narayanan and V. Shmatikov, (2009) De-anonymizing S. Networks," in Proceedings of IEEE Sym-posium on Security and Privacy, ser. SP.

Perito, D. , C. Castelluccia, M. A. K^aafar, and P. Manils, (2011) How Unique and Traceable Are Usernames?" in PETS,.

Szomszor, M. , I. Cantador, E. P. Superior, and H. Alani, (2008) Correlating user pro_les from multiple folksonomies," in In Proceedings of International Conference Hypertext (HT '08),.

Sattikar A. A. and R. V. Kulkarni., (2012) "Natural language processing for content analysis in socl networking," International Journal of Engineering Inventions, vol. 1, 6–9.

Whissell, C. (1989) "The dictionary of affect in language", Emotion: Theory, Research, and Experience, 113–131..

Weimann and Gabriel., (2010) "Terror on facebook, twitter, and youtube," The Brown Journal of World Affairs, vol. 16, 45–54,