Social Network Extraction and High Value Individual (HVI) Identification within Fused Intelligence Data

Alireza Farasat[1], Geoff Gross[1], Rakesh Nagi[2], and Alexander G. Nikolaev[1(✉)]

[1] Department of Industrial and Systems Engineering, University at Buffalo (SUNY), Buffalo, NY, USA
{afarasat,gagross,anikolae}@buffalo.edu
[2] Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Champaign, IL, USA
nagi@illinois.edu

Abstract. This paper reports on the utility of social network analysis methods in the data fusion domain. Given fused data that combines multiple intelligence reports from the same environment, social network extraction and High Value Individual (HVI) identification are of interest. The research on the feasibility of such activities may help not only in methodological developments in network science, but also, in testing and evaluation of fusion quality. This paper offers a methodology to extract a social network of individuals from fused data, captured as a Cumulative Associated Data Graph (CDG), with a supervised learning approach used for parameterizing the extraction algorithm. Ordered, centrality-based HVI lists are obtained from the CDGs constructed from the Sunni Criminal Thread and Bath'est Resurgence Threads of the SYNCOIN dataset, under various fusion system settings. The reported results shed light on the sensitivity of betweenness, closeness and degree centrality metrics to fused graph inputs and the role of HVI identification as a test-and-evaluation tool for fusion process optimization.

Keywords: Social network analysis · Data fusion · Testing and evaluation · Centrality · High value individuals

1 Introduction

Traditional data fusion has transitioned from exclusively processing hard physical sensor data to fusing soft data provided by human sources as well. In the recent multi-disciplinary university data fusion research program [10,12], Natural Language Processing together with graph analytic techniques have been applied for fusion of hard and soft information so that an analyst is able to obtain situational awareness [9] (see Figure 1). The main processing steps of this effort include (1) natural language understanding and physical sensor data processing, (2) Cumulative Data Graph (CDG) construction via association of

N. Agarwal et al. (Eds.): SBP 2015, LNCS 9021, pp. 44–54, 2015.
DOI: 10.1007/978-3-319-16268-3_5

entities and relations in attributed graphs, (3) situation assessment via graph matching identifying graphical situations of interest within the CDG, and (4) social network extraction and High Value Individuals (HVIs) identification (see [9] for details). It is hard, however, to evaluate the success of analytic approaches within fusion systems due to the upstream propagation of errors. Social net-

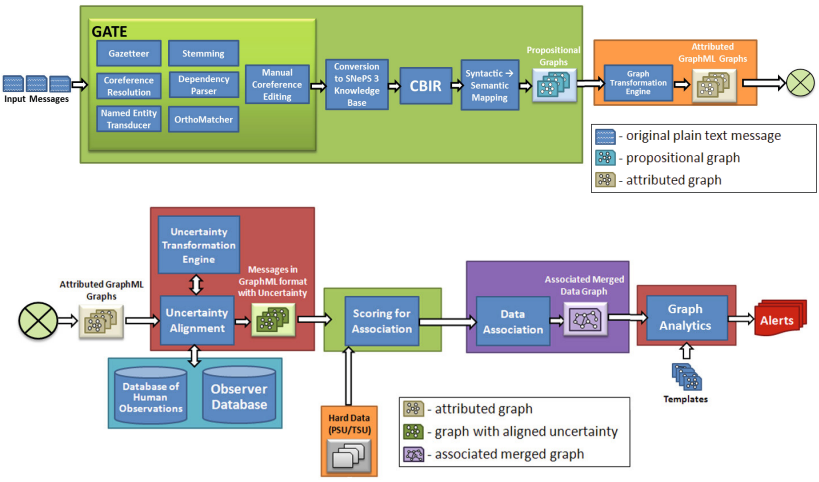

Fig. 1. Data fusion architecture in the multi-university research program

work extraction in fusion is troublesome due to complex nature of this procedure; e.g., the prediction of missing links in social network analysis (SNA) is known to be hard. When the data specifying underlying relationships is not available, inference and data mining tools are employed [21,22].

This paper reports on the use of SNA methods in the fusion domain. In evaluating a toolset for automatic processing of hard-and-soft data, one might like to reveal the relationships between the actors reported on in the data messages/signals. This paper addresses how one can assess the quality of automatic identification of HVIs, based on their structural positions in a social network [5,11,19]. The reported results shed light on the sensitivity of betweenness, closeness and degree centrality metrics to fused graph inputs and the role of HVI identification as a test-and-evaluation tool for fusion process optimization.

The paper proceeds as follows. Section 2 reviews the methodologies for extracting the social network from a CDG using a supervised learning approach for SNA of HVI identification. Section 3 provides a supervised learning approach based on Kendall Tau distance to tune our social network extractor. Section 4 discusses the results of comparison of CDG-extracted social networks (CDGSN) versus ground-truth social networks (GTSN) over multiple test datasets. Section 5 concludes the paper and discusses future research directions.

2 Social Network Extraction and HVI Modules

Social network extraction and HVI identification testing was performed within the hard-and-soft information fusion framework of the Network-based Hard+Soft Information Fusion Multidisciplinary University Research Initiative (MURI) [8–10]. To illustrate how these graph analytic tools work, the *Sunni criminal thread (SUN)* and *Bath'est Resurgence Thread (BRT)* of SYNCOIN are used [7]. Given a CDG with fused hard and soft data (e.g., locations, people, events, relationships, etc.), this section addresses the challenge of recovering the social network between person-type nodes (actors). The comparisons in recognizing HVIs are done using multiple metrics, since different metrics capture different properties of actors positions in a social network [3]. The experiments with those different metrics are conducted using supervised learning, with multiple processing modules involved in the network extraction (see Figure 2).

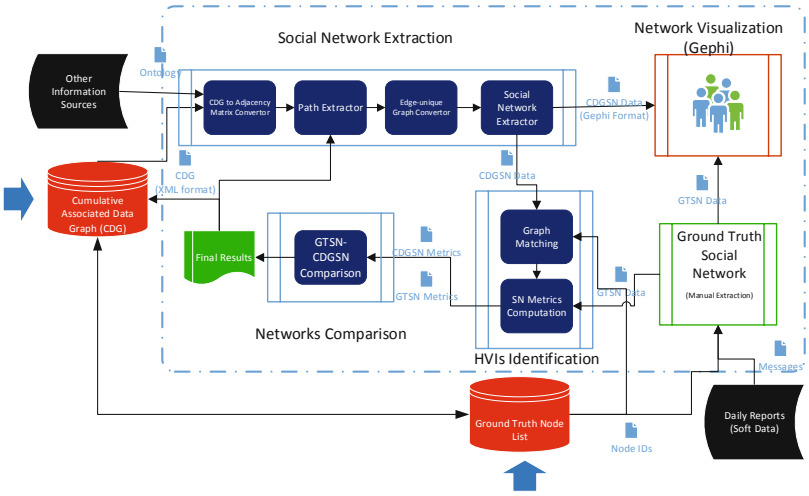

Fig. 2. Social Network Extraction and High Value Individual Identification Modules

Several methods can be used to quantify node position value in a network [1, 3, 18]; indeed, the notion of 'High Value' is application-dependent. The reported results are based on well-accepted SNA measures of centrality.

Centrality metrics quantify the prominence of actors embedded in a social network [3, 20, 23]. Based on the presumed HVI qualities and the nature of information/items exchanged between them, the betweenness, closeness or degree centrality can be adopted for HVI ranking. Degree centrality is easy to compute and reflects the number of direct connections a node has, while betweenness and closeness centralities are based on shortest paths connecting node pairs and reflect the distances between peers [4]. For the path-based centrality calculation, this study employs the very efficient algorithm of Brandes [4].

2.1 Ground Truth Social Network Extraction

For each given dataset, a corresponding Ground Truth Social Network (GTSN) is extracted by multiple analysts (those who did not participate in the dataset creation), following five steps: (1) all the raw data messages are read by each analyst, (2) all the actors mentioned, including both organization- and person-types nodes, are enumerated, (3) person-type nodes are distinguished from the complied list (as potential HVIs), (4) weights are assigned to the links connecting the persons based on the source messages, (5) the links not directly mentioned in the messages but implied are inferred (per subjective analyst judgment) by fusing the information across multiple messages. Note that controversy may arise in Steps (4-5) as the experts may not agree on the existence or the weight of a particular link. In such cases, a set of ground rules set *a-priori* by the analyst group is used for dispute resolution. To visualize the GTSN and CDGSN, an open-source software Gephi [2] is utilized (see Figure 3).

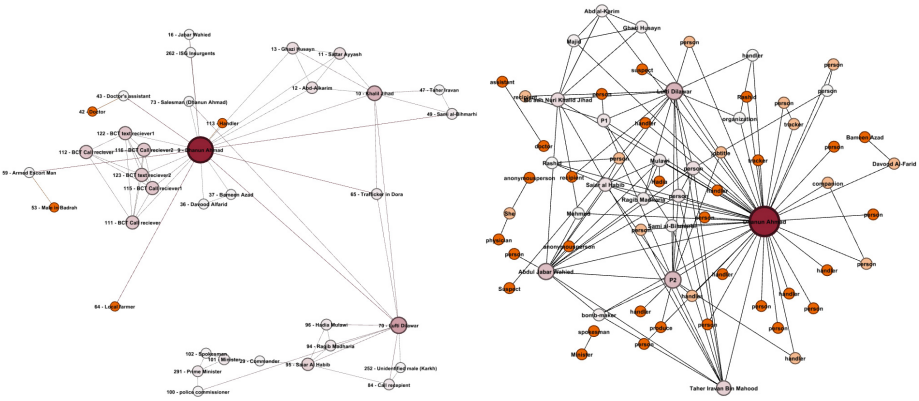

Fig. 3. Ground Truth Social Network (GTSN) (left) and Cumulative Data Graph Social Network (CDGSN) (right)

2.2 Cumulative Data Graph Social Network Extraction

The main idea employed in extracting a social network from CDG lies in traversing feasible paths between all the pairs of person-type nodes, under the feasibility and acceptance constraints. The first step is to convert a CDG XML-formatted file to an adjacency matrix of all the nodes in CDG. In extracting the feasible paths between each pair of person-type nodes, the acceptance constraints are imposed to ensure that (1) no acceptable path contains a person node, and (2) the length of an acceptable path does not exceed a pre-set threshold (T). A Depth First Search extraction procedure recovers the desired paths, with the $O(n^2T)$ time required for handling all the person node pairs.

To infer a weight of a tie between two actors (two person nodes), the number of connecting paths and the paths' lengths are taken into account. The more

connections two nodes share, the higher the weight (indicating a closer relationship between the nodes). Thus, an edge between a pair of person nodes two hops away from each other has a greater weight than an edge between a pair of person nodes who are three hops away from each other. Also, if multiple paths of the lengths smaller than the hops threshold (T) exist between a pair of nodes, then the weight of the edge between those nodes also increases. More specifically, let $w(i, j)$ denote the weight of an edge between node i and node j; P_{ij} denote a set of all paths between node i and node j; p_{ij} denote a single path such that $p_{ij} \in P_{ij}$; and $h(p_{ij})$ denote the number of hops along path p_{ij}. Then,

$$w(i,j) = \frac{1}{\sum_{p_{ij} \in P_{ij}} \frac{1}{h(p_{ij})}},$$

with $h(p_{ij}) \leq T$. Note that the inverse of this weight, $1 - w(i, j)$, is interpreted as the relationship strength, thus distinguishing strong ties (e.g., close friends) from weak ties (e.g., acquaintances) [24].

2.3 Comparison Between CDGSN and GTSN

Given two networks A and B, a simple way to assess their similarity is to count the number of changes that one has to do to transform one graph into the other (this measure is known as graph edit-distance [13]). Various edit operations have been introduced to date, including edge rotation, edge addition and subtraction, vertex addition and subtraction (if the networks do not have the same number of nodes), etc.; note that it is not obvious how to weigh these changes against one another [13].

There is a wide array of other methods proposed in the literature and exploited to measure the similarities between two given networks. Spectral analysis is used to approximate the graph edit-distance by the difference in the spectrum of eigenvalues between Laplacians of the adjacency matrices [15]. Other related research introduces p^* models (now widely known as Exponential Random Graph Models [16]), graph kernels [17] and motif analysis [14]. The p^* models and motif analysis are based on the presence of small subgraphs in the compared networks [16]. Graph kernels map graph features to points in high dimensional inner product spaces [13].

However, given the objective of HVI identification, the HVI-based measures, e.g., rankings, can be directly used to evaluate the quality of the extracted network. With the ranking of aforementioned centrality metrics used to identify HVIs, Kendall Tau distance [6] can be accordingly utilized to compare the quality of CDGSNs relative to GTSNs. Moreover, Kendall Tau can be used as a rank correlation coefficient for the evaluation of the association (agreement) between two measured lists [6] (e.g., betweenness of nodes in two different networks). The normalized Kendall Tau distance is measured as follows:

$$\tau = \frac{N_D}{\binom{N}{2}},$$

where N_D is the counts of discordance pairs and $\binom{N}{2}$ is the total number of pairs. Any pair of observations (x_i, y_i) and (x_j, y_j) is termed discordant, if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. For instance, suppose two ranked lists, $l_1 = A, C, B$ and $l_2 = B, A, C$, contain the rankings of elements A, B and C where $A = (1, 2)$ and $B = (3, 1)$ represent the rankings of A and B in l_1 and l_2, respectively. Then, A and B are discordant because A has the higher rank in l_1 and the lower rank in l_2.

3 Hop Threshold Optimization

This section describes the design of a metric for performing systemic error trail analysis and parametric optimization of the social network extraction using a supervised learning approach. Again, the main utility of an extracted social network is assumed to lie in distinguishing HVIs. The main expected utility of the extracted social network is to allow one to correctly rank HVIs for any given dataset (due to errors both in reporting and fusion, the absolute centrality values of network actors may not be reliable).

The metric utilized for minimizing the number of the HVIs misplaced in the ranked list needs to allow for the comparison of ranks of the selected centrality metrics for each node in both GTSN and CDGSN. It should be noted that CDGSN and GTSN are not necessarily expected to have the same number of nodes due to upstream processing errors. However, most or all ground truth nodes are expected to be present in both networks; in this case, the node ranks by centrality metric values can be compared using Kendall Tau distance.

Let B_i^C and B_i^G represent the betweenness values of node i in CDGSN and GTSN, respectively. Similarly, let C_i^C and C_i^G denote the closeness of node i in CDGSN and GTSN, and D_i^C and D_i^G denote degree of node i in CDGSN and GTSN ($\forall\ i \in \Omega$ where Ω is a set of common nodes in both CDGSN and GTSN). Considering the pair (B_i^C, B_i^G), the betweenness-based Kendall Tau (τ_B) is defined as follows:

$$\tau_B = \frac{|\Omega_{D_i}^B|}{\binom{|\Omega|}{2}},$$

where $\Omega_{D_i}^B$ denotes the set of discordance pairs based on node i betweenness values and $|\Omega|$ is Ω-cardinality. Similarly, the closeness-based Kendall Tau (τ_C) and degree-based Kendall Tau (τ_D) for (C_i^C, C_i^G) and (D_i^C, D_i^G), respectively, are:

$$\tau_C = \frac{|\Omega_{D_i}^C|}{\binom{|\Omega|}{2}} \quad \& \quad \tau_D = \frac{|\Omega_{D_i}^D|}{\binom{|\Omega|}{2}}.$$

Again, $\Omega_{D_i}^C$ and $\Omega_{D_i}^D$ denote the sets of discordance pairs for closeness and degree, respectively.

The training process for detecting HVIs requires one parameter to be tuned; the parameter is called the hop threshold, $T \geq 1\ T \in \mathbb{Z}$, used in the path extraction process. The goal is to find T^* as

$$T^* = \underset{T}{\operatorname{argmin}} (w_B \tau_B + w_C \tau_C + w_D \tau_D),$$

where w_B, w_C and w_D are the weights for τ_B, τ_C and τ_D, respectively. There exists no analytical solution for this optimization problem. However, a supervised learning approach can help us tune the hops threshold parameter using training datasets.

4 Training and Test Data Set Results

Two SYNCOIN dataset [7] threads were used in this test and evaluation study: the SUN thread with 114 soft messages and four hard data reports and the BRT thread with 114 soft messages. Multiple versions of CDG were generated with various upstream fusion algorithms and parameter settings utilized: fifteen CDGs for SUN and three for BRT, all in XML format.

In order to find the optimal value(s) of T, the social network extraction algorithm of Section 2 is run for $1 \leq T \leq 8$ such that the weighted sum of Kendal Tau distances is minimized. Note that this operation is computationally expensive, especially with large CDGs. Figure 4 reports the weighted sums of Kendall Tau distances for four SUN Thread CDGs. Across all the CDGs obtained under

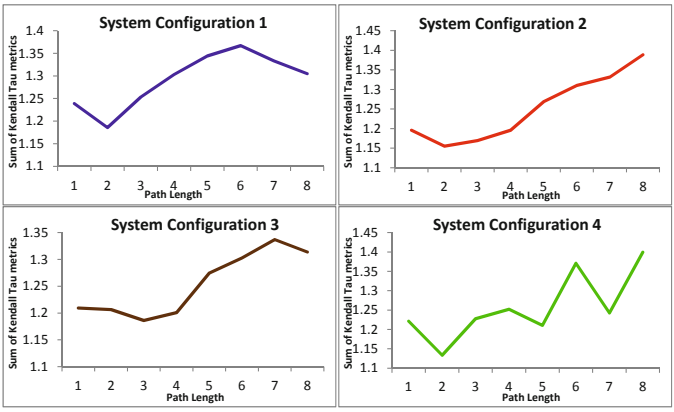

Fig. 4. Weighted Kendall Tau distance value dynamics for four SUN CDG outputs over varied values of hop threshold T

diverse fusion pipeline settings, the threshold values $T = 2$ or $T = 3$ resulted in the minimal normalized Kendall Tau distance (see Figure 4). These four configurations were chosen as exemplars demonstrating the performance under optimal data association settings for a given natural language understanding input. Naturally, the larger the value of T, the more dense the CDGSN becomes (multiple paths are found between actor pairs). Table 1 compares the top five (about top 10%) HVIs in the SUN Thread GTSN with their counterparts in CDGSN; the top 10% is a reasonable choice for the number of individuals to consider since this is reflective of the number of key actors in the SUN thread. It is found that the HVIs discovered from GTSN and CDGSN generally match. For comparing

Table 1. Top five HVIs in GTSN and CDGSN of SUN Thread dataset

Rank	Betweenness		Closeness		Degree	
	Rank in CDGSN	*Rank in GTSN*	*Rank in CDGSN*	*Rank in GTSN*	*Rank in CDGSN*	*Rank in GTSN*
1	Dhanun Ahmad	Dhanun Ahmad	Dhanun Ahmad	Dhanun Ahmad	Dhanun Ahmad	Dhanun Ahmad
2	Lufti Dilawar	Lufti Dilawar	Davood Alfarid	Lufti Dilawar	Lufti Dilawar	Lufti Dilawar
3	Trafficker in Dora	Jabar Wahied	Bameen Azad	Jabar Wahied	Khalil Jihad	Jabar Wahied
4	Khalil Jihad	Khalil Jihad	Trafficker in Dora	Davood Alfarid	Saiar Al Habib	Khalil Jihad
5	Saiar Al Habib	Saiar Al Habib	Doctor's assistant	Taher Iravan	Sattar Ayyash	Taher Iravan

the HVI ranks, betweennees centrality appears the most robust - this observation can inform the selection of the weights w_B, w_C and w_D.

With $T = 2$ or $T = 3$, it is also found that the Kendall Tau distance remains robust under different fusion settings, i.e., sub-optimal association settings as opposed to Fig 4 where optimal association settings are utilized (see Figure 5). The variability of Kendall Tau distance values at plus/minus 10% are commensurate with the error magnitudes observed in upstream data fusion stages [9].

The same conclusions about the robustness of the weighted Kendall Tau metric behavior were made in working with the BRT Thread data. Table 2 compares the top five HVIs in both GTSN and CDGSN of the BRT dataset. Again, most of top HVIs are common between BRT CDGSN and GTSN.

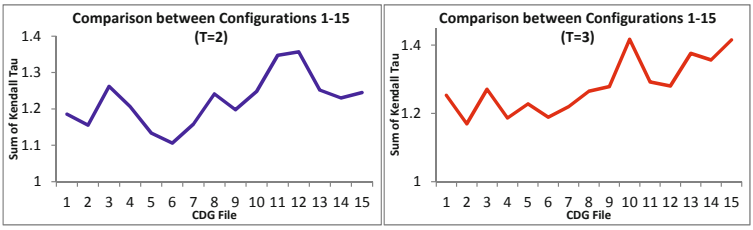

Fig. 5. Comparison of normalized Kendall Tau distance for different SUN CDG outputs

Table 2. Top five HVIs in GTSN and CDGSN of BRT dataset

Rank	Betweenness		Closeness		Degree	
	Rank in CDGSN	*Rank in GTSN*	*Rank in CDGSN*	*Rank in GTSN*	*Rank in CDGSN*	*Rank in GTSN*
1	Ali Tikriti	Khaari Elahi	Ali Tikriti	Khaari Elahi	Ali Tikriti	Khaari Elahi
2	Khaari Elahi	Ali Tikriti	Khaari Elahi	Ali Tikriti	Iraqi boy's mother	Ali Tikriti
3	MNCI Officer	MNCI Officer	Anwar Khan	Iraqi boy's mother	Khaari Elahi	Iraqi boy's mother
4	Ahmaad Hasseeb	Iraqi boy's mother	Ahmaad Hasseeb	Khalid Youssef	Anwar Khan	Khalid Youssef
5	Iraqi boy's mother	Suleiman	Khalid Youssef	MNCI Officer	Ahmaad Hasseeb	MNCI Officer

5 Conclusion

This paper extracts the underlying social networks of individuals from fused data, captured as Cumulative Associated Data Graphs (CDGs). Ordered, centrality-based HVI lists are obtained with CDGs constructed from Sunni Criminal Thread

and Bath'est Resurgence Threads of the SYNCOIN dataset, under various fusion system settings. A supervised learning approach is used to tune the social network extractor parameter - the hop threshold. Since the rank of centrality measures can guide one to recognize HVIs, the normalized Kendall Tau distance is accordingly adopted to evaluate the identified HVI rankings against those based on the ground truth data. The employed method to set parameter T is a supervised learning approach: both the training dataset and GTSN are exploited to identify the settings that provide the minimal weighted Kendall Tau distance.

The experiments with the SUN Thread and BRT Thread datasets demonstrate that HVIs can be successfully identified in CDGSN with $T \leq 3$. We recognize that error propagated from upstream processes (e.g., natural language understanding and data association) may be responsible for variations in the Kendal Tau distances. Overall, The proposed approach is useful not only in evaluating the performance of HVI identification, but also in evaluating wider, systemic data fusion performance and error propagation across fusion processes. We plan to extend this analysis to the wider fusion system in future work. It is observed that betweenness centrality appears more robust than other centrality measures in HVI ranking.

Future work will focus on path semantics in HVI identification recognizing link types: e.g., a familial and organizational link is stronger than a locational link. Fusing alternative paths to arrive at "composite weights" is also promising.

Acknowledgments. This work has been supported by a Multidisciplinary University Research Initiative (MURI) grant (Number W911NF-09-1-0392) for Unified Research on Network-based Hard/Soft Information Fusion, issued by the US Army Research Office (ARO) under the program management of Dr. John Lavery.

References

1. Arulselvan, A., Commander, C.W., Elefteriadou, L., Pardalos, P.M.: Detecting critical nodes in sparse graphs. Computers & Operations Research **36**(7), 2193–2200 (2009)
2. Bastian, M., Heymann, S., Jacomy, M., et al.: Gephi: an open source software for exploring and manipulating networks. ICWSM **8**, 361–362 (2009)
3. Borgatti, S.P.: Identifying sets of key players in a social network. Computational & Mathematical Organization Theory **12**(1), 21–34 (2006)
4. Brandes, U.: A faster algorithm for betweenness centrality. Journal of Mathematical Sociology **25**(2), 163–177 (2001)

5. Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C., Zhou, T.: Identifying influential nodes in complex networks. Physica a: Statistical Mechanics and its Applications **391**(4), 1777–1787 (2012)
6. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top k lists. SIAM Journal on Discrete Mathematics **17**(1), 134–160 (2003)
7. Graham, J.L., Hall, D.L., Rimland, J.: A synthetic dataset for evaluating soft and hard fusion algorithms. SPIE Defense, Security, and Sensing International Society for Optics and Photonics, pp. 80620F–80620F (2011)
8. Gross, G., Nagi, R., Sambhoos, K.: A fuzzy graph matching approach in intelligence analysis and maintenance of continuous situational awareness. Information Fusion **18**, 43–61 (2014)
9. Gross, G.A., Khopkar, S., Nagi, R., Sambhoos, K., et al.: Data association and graph analytical processing of hard and soft intelligence data. In: 2013 16th International Conference on Information Fusion (FUSION), pp. 404–411. IEEE (2013)
10. Gross, G.A., Nagi, R., Sambhoos, K., Schlegel, D.R., Shapiro, S.C., Tauer, G.: Towards hard+soft data fusion: Processing architecture and implementation for the joint fusion and analysis of hard and soft intelligence data. In: 2012 15th International Conference on Information Fusion (FUSION), pp. 955–962. IEEE (2012)
11. Kiss, C., Bichler, M.: Identification of influencers - measuring influence in customer networks. Decision Support Systems **46**(1), 233–253 (2008)
12. Llinas, J., Nagi, R., Hall, D., Lavery, J.: A multi-disciplinary university research initiative in hard and soft information fusion: Overview, research strategies and initial results. In: 2010 13th Conference on Information Fusion (FUSION), pp. 1–7. IEEE (2010)
13. Macindoe, O., Richards, W.: Graph comparison using fine structure analysis. In: 2010 IEEE Second International Conference on Social Computing (SocialCom), pp. 193–200. IEEE (2010)
14. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. Science **298**(5594), 824–827 (2002)
15. Mitchell Peabody, Finding groups of graphs in databases, Ph.D. thesis, Citeseer (2002)
16. Robins, G., Pattison, P., Kalish, Y., Lusher, D.: An introduction to exponential random graph p-star models for social networks. Social Networks **29**(2), 173–191 (2007)
17. Shervashidze, N., Borgwardt, K.M.: Fast subtree kernels on graphs. Advances in Neural Information Processing Systems, 1660–1668 (2009)
18. Shetty, J., Adibi, J.: Discovering important nodes through graph entropy the case of enron email database. In: Proceedings of the 3rd International Workshop on Link Discovery, pp. 74–81. ACM (2005)
19. Rama Suri, N., Narahari, Y.: Determining the top-k nodes in social networks using the shapley value. In: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, vol. 3. International Foundation for Autonomous Agents and Multiagent Systems, pp. 1509–1512 (2008)
20. Tang, J., Musolesi, M., Mascolo, C., Latora, V., Nicosia, V.: Analysing information flows and key mediators through temporal centrality metrics. In: Proceedings of the 3rd Workshop on Social Network Systems, p. 3. ACM (2010)

21. Thelwall, M., Wilkinson, D., Uppal, S.: Data mining emotion in social network communication: Gender differences in myspace. Journal of the American Society for Information Science and Technology **61**(1), 190–199 (2010)
22. van der Aalst, W.M.P., Song, M.S.: Mining Social Networks: Uncovering Interaction Patterns in Business Processes. In: Desel, J., Pernici, B., Weske, M. (eds.) BPM 2004. LNCS, vol. 3080, pp. 244–260. Springer, Heidelberg (2004)
23. Varlamis, I., Eirinaki, M., Louta, M.: A study on social network metrics and their application in trust networks. In: 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 168–175. IEEE (2010)
24. Xiang, R., Neville, J., Rogati, M.: Modeling relationship strength in online social networks. In: Proceedings of the 19th International Conference on World Wide Web, pp. 981–990. ACM (2010)