Visualizing Social Network Data: a comparative study of Asian-American student conferences

Roberto Palmieri
Dept. Mechanical, Energy & Management Engineering
University of Calabria
Rende, Italy
e-mail: roberto.palmieri@unical.it

Carlo Giglio
Dept. Mechanical, Energy & Management Engineering
University of Calabria
Rende, Italy
e-mail: carlo.giglio@unical.it

***Abstract*—In this paper we propose a comparative study of social network data related to three Asian-American student conferences: Taiwan-America Student Conference (TASC), Japan-America Student Conference (JASC) and Korea-America Student Conference (KASC). Such a study is built on the literature review of existing visualization methods and is based on the adoption of open source and freely available tools for Social Network Analysis (SNA). The main aim of this work is that of analyzing and comparing interaction patterns and sub-networks dynamics of attending students emerging from the collected data. In particular, data have been extracted in wide temporal horizons starting 30 days before and finishing 30 days after the application deadlines.**

Keywords-Social Network Analysis; Data Visualization; Visualization Methods; Case Study; Asian-American Student Conferences.

I. INTRODUCTION

Today's student conferences are often organized with the main aim of helping young people to establish stable relationships by a social and professional standpoint. Moreover, this kind of events represents also an effective channel for knowledge exchange among new generations coming from different countries. The idea behind many of these conferences is that of providing young people with a shared vision of a peaceful future and of encouraging them to collaborate and share their ideas and knowledge. As a matter of fact, knowledge sharing is recognized as a key factor in order to develop innovative solutions for a better world [1] [2] [3] [4] [8].

By the attendees' standpoint, gaining new experiences and sharing knowledge tends to ensure a higher performance in social and working contexts since participants are more likely to develop significant competitive advantages [5] [6].

Nonetheless the abundance of visualization methods in literature may generate some issues when dealing with the choice of the right technique for analyzing specific case studies [24]. Therefore, choosing the right methodology is directly linked to the level of understanding of the problem at hand. In light of the importance of such techniques, the research community continuously pays attention to many visualization-related topics [24]. In fact, based on the way data and information are showed, the process of elaboration and analysis may change relevantly, since different factors and issues may arise depending on the specific visualization strategy adopted. Visualization-related key issues affecting elaboration and analysis of social network data concern, among others, with the design of the visualization process, the visualization methods clustering process, and possible combinations of different methods.

The development of a number of visualization solutions is due to the even bigger volume of data and information available in different contexts [25]. However, it is important to notice how such visualization tools are mainly connected to knowledge extraction activities from available data more than to data collection processes. In fact, visualization tools play an influence on the approach chosen by data analyst and scientist, since they address the setting up of the data analysis strategy and processes [25]. In this context, adopting advanced visual tools aligned with individuals' analysis capabilities is a key point in order to ensure the best interpretation of available data [26] [27].

We propose a comparative study of three student conferences through visualization tools in order to analyze interaction patterns and sub-networks dynamics of participants.

In particular, such an analysis may help identifying in advance hot topics dealt with during the conferences and potential sub-groups of students having some interests in common.

In light of the role played by such tools in terms of understanding of many phenomena in a number of fields, in Section II we introduce some theoretical issues and definitions about visualization process, methods, clustering, and corresponding factors. In Section III, the methodological aspects are detailed - e. g. research method, context-dependent methodological aspects, data collection and extraction. In Section IV, results and findings of the comparative study are presented. Section V concludes by addressing possible applications of this study in other fields, future research efforts and possible limitations.

978-1-4673-8103-1/15 $31.00 © 2015 IEEE
DOI 10.1109/FiCloud.2015.100

II. Theory About Visualization Methods And Related Research

This section provides an overview of visualization methods and processes, and corresponding factors. In particular, theoretical issues and definitions from related research are discussed.

A. Visualization Process and Corresponding Factors

According to some authors [24] [28], visualization is intended to be the representation of a phenomenon of interest through tools rendering it similar to visual perception. Such tools are computer-supported and enhance human capabilities in terms of knowledge acquisition [29] [30] by means of a graphical description, which is richer and more intuitive at the same time and ensure analysts to explore, search, communicate, show in a visual way [31]. Moreover, it allows visualization experts to show, rank and cluster data under different perspectives and to adapt them for analysis of different cognitive processes in many contexts [32] [33].

The existing literature about the design of the process of visualization provides a six-step model [34]. Such steps are: mapping, selection, presentation, interactivity, usability, and evaluation, as detailed in the following:

1) Mapping deals with the way data and information are encoded and rendered visually. Encoding data and information correctly is a key issue in order to ensure the alignment between real-world features and their corresponding representation by means of visualization tools. An adequate algorithm is thus required in order to depict such features in visual form.
2) Selection is recognized to be the most important step, since in it data scientists and analysts decide, among all available data, which ones have to be thrown away and which ones are useful. Such a decision also depends on the specific task data are going to be used for. Wrong selections may lead to project or process failures.
3) Presentation concerns with the way data and information are provided to the target audience. Therefore, it is very important to organize data and information properly in order to ensure that they are intelligible.
4) Interactivity is geared to provide users with tools and features such that data and information can be investigated, scrutinized and reworked.
5) Usability and accessibility are very important, since they introduce in this six-step model the concept of "human factor". Visualized data and information should take into account the accessibility need of special categories of users and ensure easy of use of people.
6) Evaluation is directly related to measuring the achievement of expected results. In particular, two key issues are considered in this step: the degree of effectiveness and the achievement of visualization goals. Evaluation can be performed empirically by means of questionnaires, interviews, focus groups, and controlled experiments, while analytic evaluation may include cognitive walkthrough and expert reviews [29].

B. Visualization Methods

In the following a brief introduction to data visualization methods is provided. Moreover, such methods are briefly described [24]:

a) Table is a very intelligible method of data representation with a widespread format containing set of variables and values. Despite the general format is well structured, table is also a flexible tool, which allows to develop several variants based on the specific research goals.
b) Pie Chart is a disk-shaped or round-shaped visualization tool containing several slices. Each slice is generally associated with a variable and its corresponding percentage value. Advanced variants of such a tool allow to manage also hierarchical data.
c) Bar Chart is a widespread tool for visual representation of data. Likewise in the other cases, also the bar chart format may be modified in order to achieve specific research-related goals.
d) Histogram is a visual tool dealing with several clusters of variables. Objects are classified into such categories of variables based on the statistical elaboration of available data.
e) Line Chart is one of the most adopted visualization tools consisting of a number of points connected to each other.
f) Area Chart is a graph for representing data in bounded area.
g) Scatter Plot is a data visualization tool geared to show points distributed in Cartesian coordinate.
h) Bubble Chart is a kind of Scatter Plot with points have Cartesian coordinates, but are also associated with an additional value that is the diameter of the bubble point.

Multiple Data Series is a combined visualization tool including the above mentioned methods.

III. Methodological Section

This section deals with the methodological aspects of this research.

In a conference setting - especially those geared to students, which are generally lacking in experience if compared to professionals and academicians – knowledge and idea sharing is part of the dynamics planned by event managers and emerging from disputes among participants [9].

Nonetheless, student conferences are characterized by cultural differences – which are expected to be significant when considering Asian-American exchange contexts –, possible problems to develop a shared vision, and communication difficulties among two sided-conference participants. Therefore, triggers planned by conference managers may prove to be useless if attendees are not inclined to make the proper efforts in order to deepen hot topics during the event. Moreover, event managers affect attendees' behavior since their triggering strategies and plans may fail also due to the specific context rules of the conference [7].

The event setting affects the way people result to be engaged in the conference activities, but such an influence is exerted also by innovative solutions – i. e. technologies and tools -, which may enhance participants engagement. In this context, online communities are recognized to be the best setting for attendees in order to clearly state what their perceptions and ideas are [7].

Given the importance of such technologies and tools, event organizers adopted them with the main aim of fostering collaborative behaviors among participants, knowledge sharing, and the development of a shared vision among students with relevant cultural differences [10]. Social media helps triggering in advance some conference dynamics planned by the organizers. Online communities provide data about hot topics, hence, addressing event dynamics before the start of the conference.

A. Research Method

Participants' interactions during student conferences can be analyzed by obtaining data from specific online communities, hence, assuming that such communities exist and may provide interesting and reliable data about Asian-American student networks [11] [13].

This paper is based on the data visualization and analysis [15] of three conference-related communities on Facebook, namely the Taiwan-America Student Conference (TASC), the Japan-America Student Conference (JASC) and the Korea-America Student Conference (KASC). The data extraction process concerns with a wide temporal horizons starting 30 days before and finishing 30 days after the application deadlines.

This paper is built on a research method, which I fully developed on the basis of the data science research approach [12] and all the data collection and extraction methods and tools adopted in this study have been applied on online sources [14].

B. Context-dependent methodological issues related to the comparative study

The case studies analyzed in this work are those related to three Facebook communities of Asian-American student conferences (TASC, JASC and KASC), which will take place in July-August 2015. For each student conference, the event setting is triggered by the corresponding organizers and participants on the proper Facebook community. The three events are annually organized. Despite social media concerns with only a quota of overall attendees' interactions, the three conferences are attended by young and highly connected students willing to share ideas, knowledge and experiences, and to meet new people and learn new things. Hence, the three Facebook communities – i. e. TASC, JASC and KASC, which count on 2,439 users, 1,481 users and 654 users, respectively - and the data collection and extraction methods are able to provide a huge and representative amount of data to be analyzed.

C. Data collection and extraction

The collection and extraction methods adopted for this research included the NetVizz app v1.2 [16] in order to obtain page data by logging in into Facebook with a generic user account. NetVizz is able to provide academicians with data from Facebook and it has been developed by Professor Bernhard Rieder. NetVizz is a reliable tool, which provides data compatible with many visualization tools [16]. Gephi [17] is the open source software adopted for data visualization and is compatible with data from NetVizz. Gephi includes a rich set of solutions for data visualization and analysis – e. g. algorithms and filters, personalization options, flexibility, scalability, WYSIWYG and user-friendly software. In Table 1 data for the NetVizz queries are shown in order to ensure transparency and reproducibility of this comparative study.

TABLE I. DATA FOR THE NETVIZZ QUERIES.

Student conference	**Facebook ID**	**Data collection and extraction (day/month/year)**		**Likes**
		From	***To***	
TASC	225845154251232	29/01/15	30/03/15	2,439
KASC	333942935541	04/10/14	01/02/15	654
JASC	114135335328456	01/12/14	30/01/15	1,481

IV. RESULTS AND FINDINGS

Gephi counts 122 nodes and 281 directed edges in the resulting graph of TASC, while it associates 143 nodes and 190 directed edges to JASC, and 40 nodes and 52 directed edges to KASC.

First, we measured graph density in order to understand how graphs are close to complete. Graph density of JASC is equal to 0,009, thus showing a low level of connection among its 1,481 members. Graph density of TASC is equal to 0,019, while KASC proves to be a density level of 0,033.Hence, the highest density among the three conference-related Facebook communities is that of KASC, which has also the lowest number of likes.

Low levels of graph density have a negative influence on event organizers' triggering activities, that is a graph with lower density is likely to show less interactions among participants. However, other interesting data about the quality of users interactions – which are not available in the dataset extracted from Facebook, due to recent changes in its privacy policy, as detailed in Section V – should be taken into account in future research efforts.

Given the above results, we started analyzing possible strongly and weakly connected components (hereinafter SCC and WCC, respectively) in the three graphs [18]. Figures 1, 2 and 3 shows the above mentioned results.

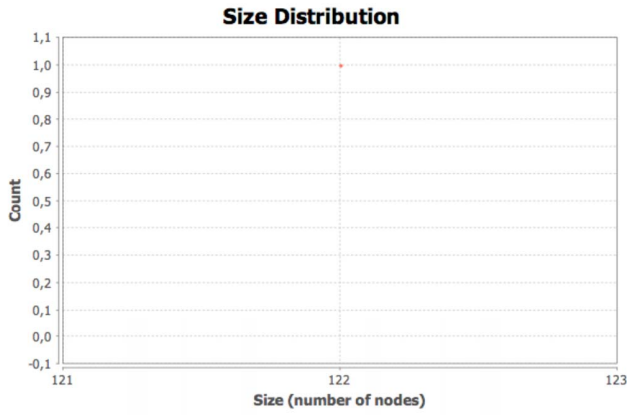

Figure 1. Strongly and weakly connected components for TASC.

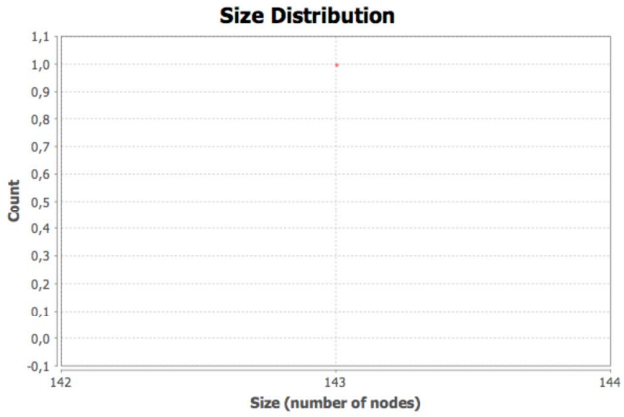

Figure 2. Strongly and weakly connected components for JASC.

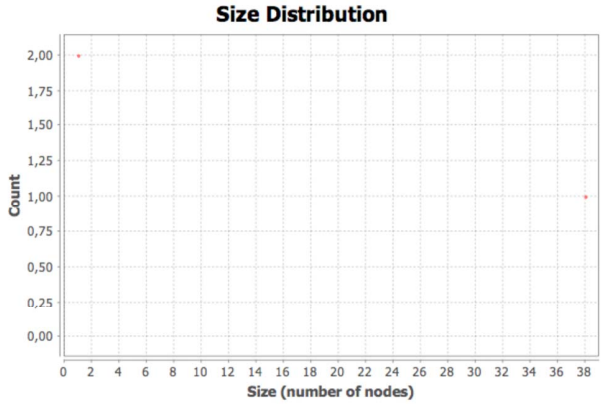

Figure 3. Strongly and weakly connected components for KASC.

For TASC, we obtained a measure of 1 weakly connected components and of 122 strongly connected ones. For JASC, weakly connected components are 1, while strongly connected components are 143. For KASC, weakly connected components are 3, while strongly connected components are 40.

Combining the graph density analysis and the connected components measures, results show how some sub-networks tend to live the pre-conference online experience separated from each other, though most of the considered sub-groups prove to be strongly tied and to have developed solid intra-component relationships. Results also suggest that users interactions in each Facebook page are not restricted to marginal and elite groups, but they are quite common in the overall community of users. Privacy settings do not make it possible to know whether an online sub-group mirrors user aggregation dynamics in the real world or it represents a different kind of group built on non-virtual relationships.

Sub-networks in this work can be studied also by adopting modularity measures (Figures 4, 5 and 6). Modularity concerns with community detection algorithms developed for this specific analysis. The above analysis is realized by using standard parameters and resolution values [19] [20]. Modularity and modularity with resolution tend to be very similar for two communities out of three that is KASC and JASC (>0,500), while they are slightly lower for TASC (=0,344). Moreover, TASC shows the highest number of communities – i. e. 9 –, while KASC and JASC provided lower values – i. e. 7 and 5, respectively.

Obviously, such values have been calculated on the basis of the same parameters. In light of this, data about the three conferences are homogeneous.

Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.0

Results:

Modularity: 0,344
Modularity with resolution: 0,344
Number of Communities: 9

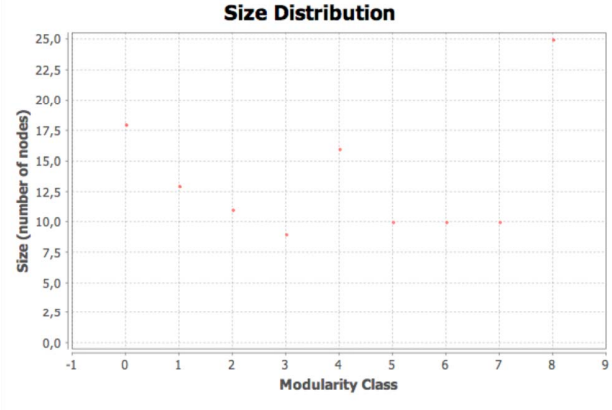

Figure 4. Modularity for TASC.

Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.0

Results:

Modularity: 0,501
Modularity with resolution: 0,501
Number of Communities: 5

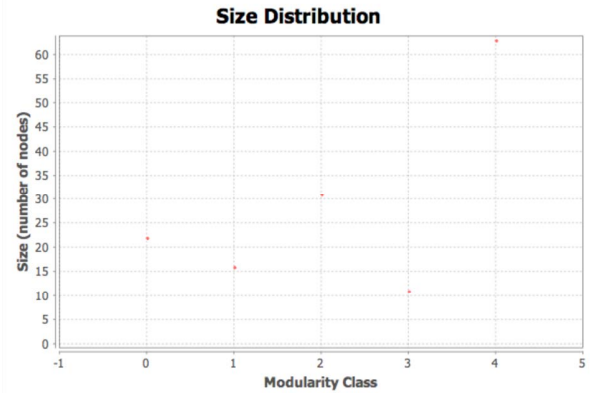

Figure 5. Modularity for JASC.

Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.0

Results:

Modularity: 0,539
Modularity with resolution: 0,539
Number of Communities: 7

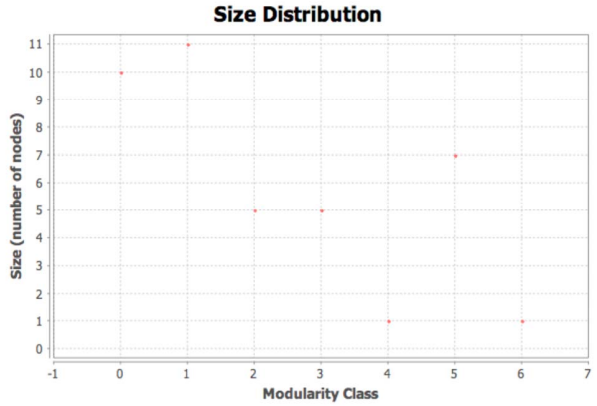

Figure 6. Modularity for KASC.

Shortest paths (SP) are 52 in KASC, network diameter (ND) and average path length (APL) are equal to 1. SP are 190 in JASC, ND and APL are equal to 1. SP are 281 in TASC, ND and APL are equal to 1. HITS and centrality measures are shown in Figures 7, 8 and 9 [21] [22] and PageRank in Figures 10, 11 and 12 [23].

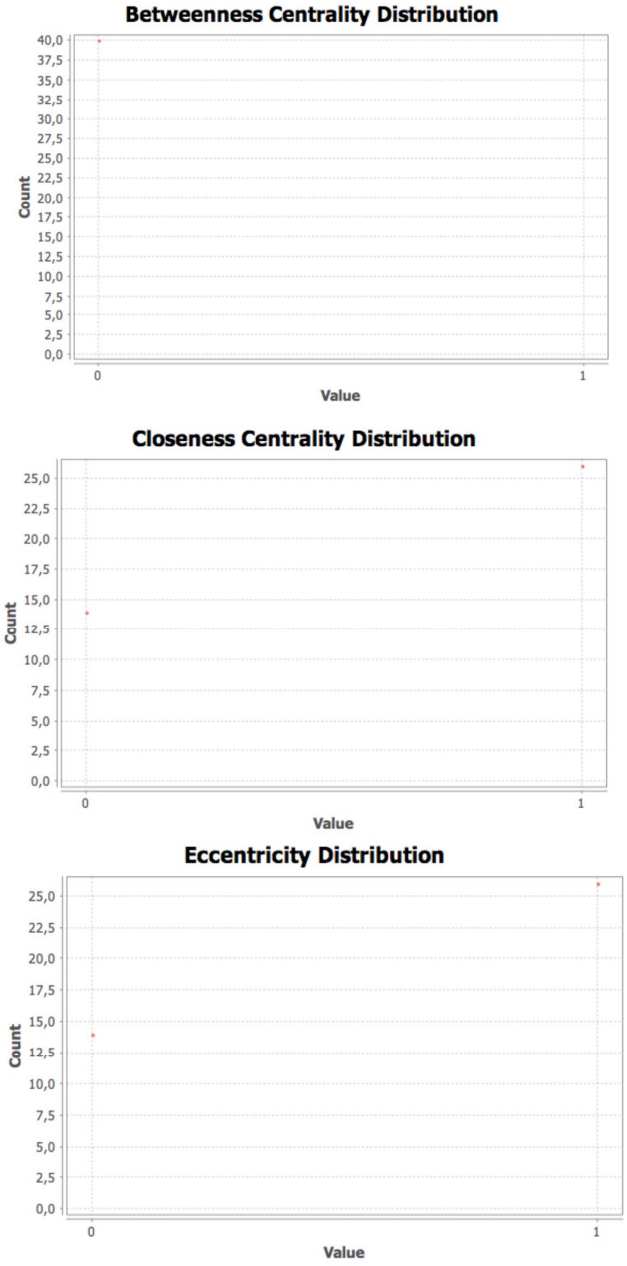

Figure 7. HITS and centrality report for KASC.

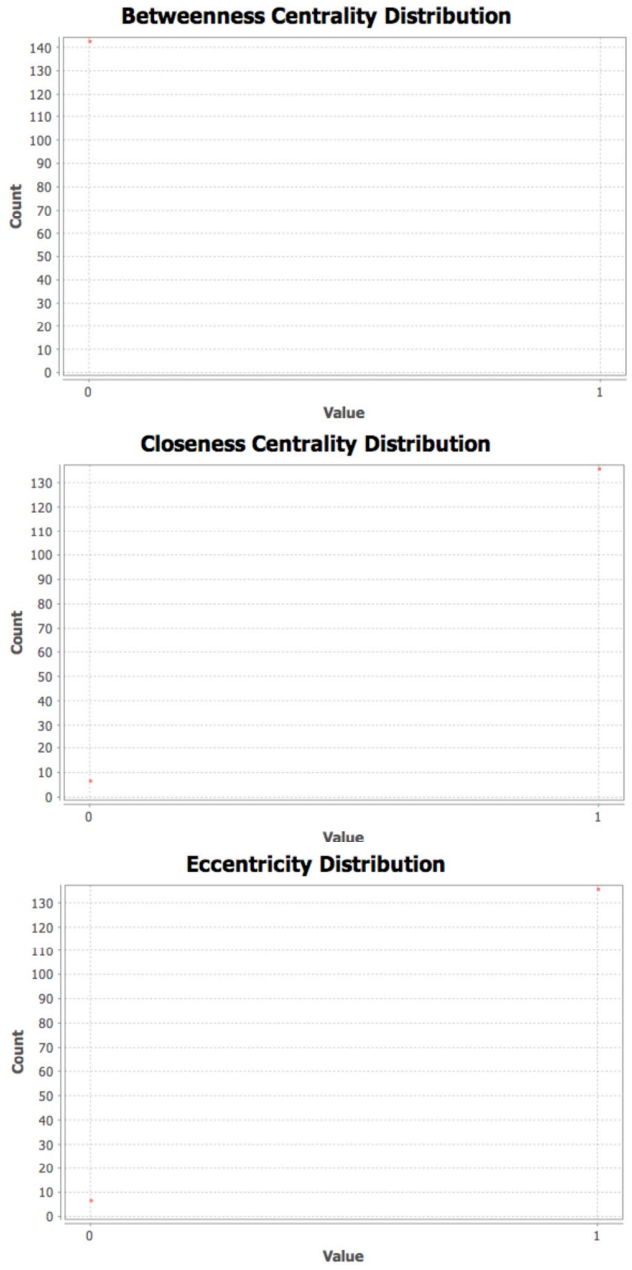

Figure 8. HITS and centrality report for JASC.

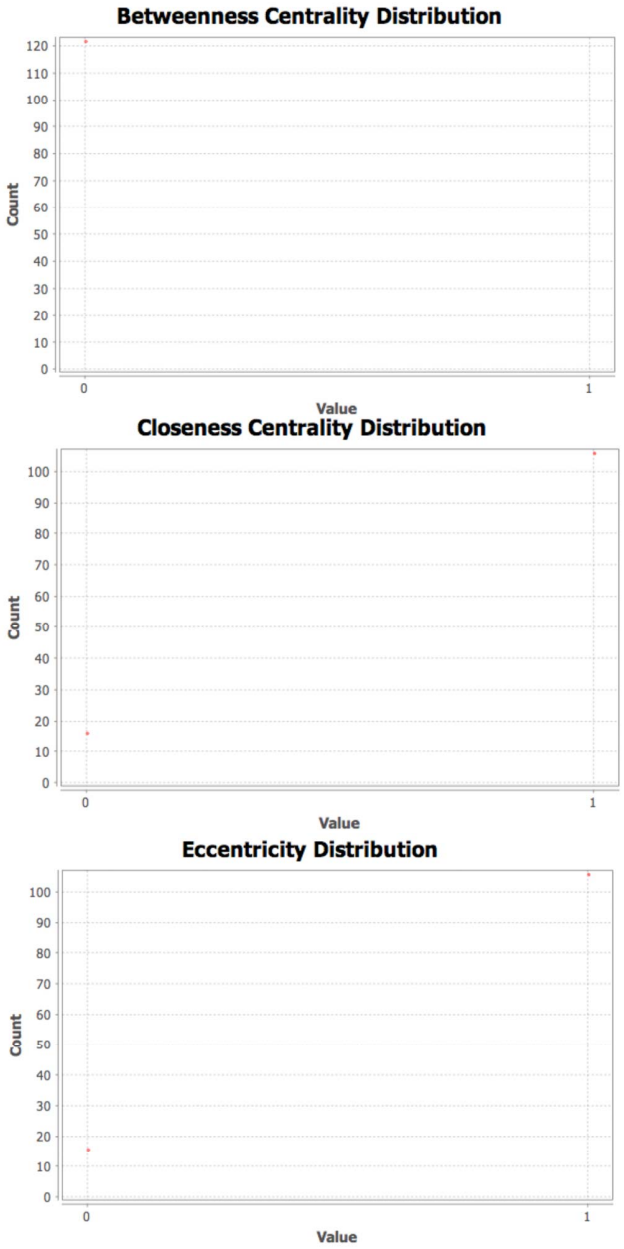

Figure 9. HITS and centrality report for TASC.

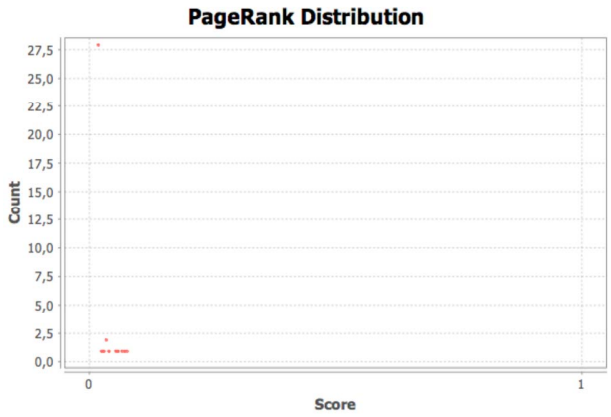

Figure 10. PageRank distribution for KASC.

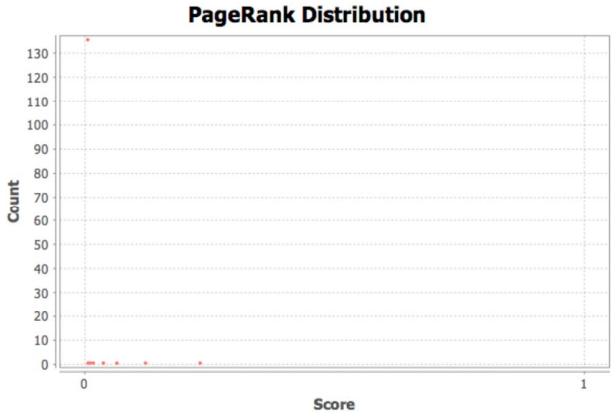

Figure 11. PageRank distribution for JASC.

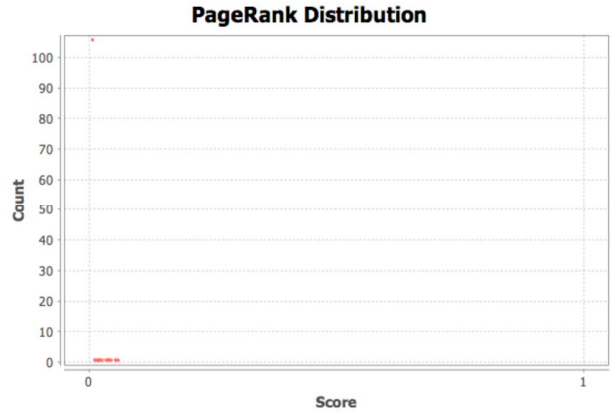

Figure 12. PageRank distribution for TASC.

Ultimately, the overall analysis of the network tends to confirm that, despite the existence of disconnected components, there is a high degree of connectedness within most components.

In conclusion, data collected and extracted and the corresponding results and findings are visually represented in Table 2. Discussion and conclusions in the final section are based on such data representation framework.

TABLE II. DATA VISUALIZATION FROM GEPHI

Data/Conference	KASC	JASC	TASC
Size distribution (# of nodes)	40	143	122
# of directed edges	52	190	281
Graph density	0.033	0.009	0.019
Weakly Connected Components (WCC)	3	1	1
Strongly Connected Components (SCC)	40	143	122
Modularity	0.539	0.501	0.344
Modularity with resolution	0.539	0.501	0.344
Number of communities	7	5	9
Shortest paths (SP)	52	190	281
Network diameter (ND)	1	1	1
Average path length (APL)	1	1	1
Betweenness Centrality Distribution (BCD) count range for values = 0	[40-ε; 40+ε]	[140; 150]	[120; 130]
Betweenness Centrality Distribution (BCD) count range for values = 1	[0; 0+ε]	[0; 0+ε]	[0; 0+ε]
Closeness Centrality Distribution (CCD) count range for values = 0	[12,5; 15,0]	[0; 10]	[10; 20]
Closeness Centrality Distribution (CCD) count range for values = 1	[25; 27,5	[130; 140]	[100; 110]
Eccentricity Distribution (ED) count range for values = 0	[12,5; 15,0]	[0; 10]	[10; 20]
Eccentricity Distribution (ED) count range for values = 1	[25; 27,5	[130; 140]	[100; 110]
PageRank Distribution count range for values tending to 0	[40-ε; 40+ε]	[140; 150]	[120; 130]
PageRank Distribution count range for values tending to 1	[0; 0+ε]	[0; 0+ε]	[0; 0+ε]

V. Discussion And Conclusions

The attributes in the dataset extracted from Facebook by means of the NetVizz app do not include useful data for the identification of emerging hot topics. This is due to the very recent changes in the privacy policy of Facebook, which currently makes it very difficult to obtain reliable and significant amounts of data about users' privacy-related fields. Hence, possible hot topics emerging from pre-conference triggering efforts planned by the event managers, may seem to be a failure even if they are successful. However, the research design proved to be useful since it allowed to analyze and to identify interactions among attendees in the pre-conference period, despite the above mentioned difficulty to retrieve specific topics-related data

attributes from Facebook. This reveals how the approach proposed in this study may lead to overcome some obstacles coming from relevant exogenous factors – e. g. changes in external processes related to Facebook privacy policy. Moreover, this proves how the research methodology keeps unchanged its reliability also when the expected quality of available data changes.

The comparative study makes it possible to observe and to analyze users interactions, thus allowing to predict both hot topics and human behavior in similar contexts. Given the exploratory peculiarity of the research at hand, it is geared to prepare the way for field scholars in order to ensure more accurate research efforts about the quality of attendees' interactions in the future. Hence, this work proves the importance of deepening the prediction of users' behavior and conference hot topics through new visualization methods. Moreover, further studies can be conducted in order to identify more advanced solutions, which could be adopted also for different research task and in different contexts.

By a methodological perspective, a novel and emerging approach is adopted. It is linked to the most recent data visualization tools and applied to a specific field – i. e. the comparative study of Asian-American student conferences.

Ultimately, this work may act as a key, consolidate basis for developing innovative sub-networks detection tools and techniques, and topics-related prediction algorithms.

REFERENCES

[1] Yusuf, S. (2009) 'From creativity to innovation', Technology in Society, Elsevier.

[2] de Castro, E. A., Rodrigues, C., Esteves, C. and da Rosa Pires, A. (2000) 'The triple helix model as a motor for the creative use of telematics', Research Policy. Elsevier.

[3] Burton-Jones, A. (2001) 'The knowledge supply model: a framework for developing education and training in the new economy', Education and Training.

[4] Iammarino, S. (2005) 'An evolutionary integrated view of Regional Systems of Innovation: concepts, measures and historical perspectives', European planning studies.

[5] Di Pietro, W. and Anoruo, E. (2006) 'Creativity, innovation, and export performance', Journal of Policy Modeling, Elsevier.

[6] Takeuchi, H. (2006) 'The new dynamism of the knowledge creating company. In Japan, moving toward a more advanced knowledge economy: advanced knowledge-creating companies'. Volume 2. Eds. H. Takeuchi & T. Shibata. Washington: World Bank. 1-9.

[7] Aramo-Immonen H., Jussila J. And Huhtamäki J. (2014) 'Visualizing informal learning behavior from conference participants Twitter data', TEEM '14 Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality, Pages 603-610.

[8] Roberto Palmieri , Carlo Giglio, (2014) "Seeking the stakeholder-oriented value of innovation: a CKI perspective", Measuring Business Excellence, Vol. 18 Iss: 1, pp.35 – 44.

[9] Engeström, Y. 2000. Activity theory as a framework for analyzing and redesigning work. Ergonomics. 43, 7 (2000), 960–974.

[10] Jussila, J., Huhtamäki, J., Kärkkäinen, H. and Still, K. 2013. Information visualization of Twitter data for co-organizing conferences. Proceedings of the 17th International Academic MindTrek Conference: Making Sense of Converging Media (Tampere, 2013).

[11] Card, S.K., Mackinlay, J.D. and Shneiderman, B. 1999. Readings in information visualization: using vision to think. Morgan Kaufmann Pub.

[12] Hey, A.J., Tansley, S. and Tolle, K.M. 2009. The fourth paradigm: data-intensive scientific discovery. (2009).

[13] Benbasat, I., Goldstein, D.K. and Mead, M. 1987. The case research strategy in studies of information systems. MIS Quarterly. (1987), 369–386.

[14] Davenport, T. 2014. Big data at work: dispelling the myths, uncovering the opportunities. Harvard Business Review Press.

[15] Ware, C. 2004. Information Visualization: Perception for Design. Elsevier.

[16] B. Rieder (2013) 'Studying Facebook via data extraction: the Netvizz application', In WebSci '13 Proceedings of the 5th Annual ACM Web Science Conference (pp. 346-355). New York: ACM.

[17] Bastian M., Heymann S. and Jacomy M. (2009) 'Gephi: an open source software for exploring and manipulating networks', Proceedings of the Third International ICWSM Conference.

[18] Robert Tarjan, Depth-first search and linear graph algorithms, in SIAM Journal on Computing 1 (2): 146–160 (1972).

[19] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre (2008) Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment (10), P1000.

[20] R. Lambiotte, J.-C. Delvenne, M. Barahona (2009) Laplacian dynamics and multiscale modular structure in networks.

[21] Ulrik Brandes (2001) A faster algorithm for betweenness centrality, in Journal of Mathematical Sociology 25(2):163-177.

[22] Jon M. Kleinberg (1999) Authoritative sources in a hyperlinked environment, in Journal of the ACM 46 (5): 604–632.

[23] Sergey Brin, Lawrence Page (1998) The anatomy of a large-scale hypertextual web search engine, in Proceedings of the seventh International Conference on the World Wide Web (WWW1998):107-117.

[24] Khan M. and Khan S. S. (2011), Data and information visualization methods, and interactive mechanisms: a survey, International Journal of Computer Applications, Volume 34– No.1.

[25] Chris North, "Information visualization", Center for Human-Computer Interaction, Department of Computer Science Virginia Polytechnic Institute and State University Blacksburg, VA 24061 USA.

[26] Ware, C. (2004). "Information visualization: perception for design", Morgan Kaufmann.

[27] Spence, R. (2001). "Information visualization", AddisonWesley.

[28] S. Card, J. MacKinlay, and B. Shneiderman, (1998). "Readings in information visualization: using vision to think". Morgan Kaufmann.

[29] Alfredo R. Teyseyre and Marcelo R. Campo, (2009). "An overview of 3D software visualization", IEEE Transactions on Visualization and Computer Graphics, vol.15, No.1.

[30] WebSter Dictionary (2011), in Khan M. and Khan S. S., Data and information visualization methods, and interactive mechanisms: a survey, International Journal of Computer Applications, Volume 34– No.1, November 2011.

[31] E.R. Tufte, (1997). "Visual explanations: images and quantities, evidence and narrative", Graphics Press, 1997.

[32] Gerald J. Kowalski and Mark A. Maybury. Information storage and retrieval system, theory and implementation. Second Edition.

[33] D.M. Butler, J.C. Almond, R.D. Bergeron, K.W. Brodlie , and A.B. Haber, (1993). "Visualization reference models", Proc. Fourth IEEE Conf. Visualization, G.M. Nielson and D. Bergeron, eds., pp. 337-342.

[34] L. Chittaro, (2006). "Visualizing information on mobile devices", ACM Computer, v.39 n.3, p.40-45.