ICIT 2015 The 7th International Conference on Information Technology
doi:10.15849/icit.2015.0082

Exploring Domain Interrelations in Freebase Schema Using Modularity-Based Community Detection

Mahmoud Elbattah
College of Engineering and Informatics, National University of Ireland
m.elbattah1@nuigalway.ie

Mohamed Roshdy
Faculty of Computer and Information Sciences Ain Shams University, Cairo, Egypt
mroushdy@cis.asu.edu.eg

Mostafa Aref
Faculty of Computer and Information Sciences Ain Shams University, Cairo, Egypt
aref_99@yahoo.com

Abdel-Badeh Salem
Faculty of Computer and Information Sciences Ain Shams University, Cairo, Egypt
abmsalem@yahoo.com

***Abstract*— Freebase is intended to be an important component of the Linked Open Data (LOD). The paper presents a graph-driven methodology for the analysis and visualisation of Freebase complex schema. First, the methodology utilises Freebase schema types, "Included Type" relationships and "Instance Count" properties to construct a directed weighted graph schema. Second, the schema graph is employed to conduct modularity-based analysis in order to detect communities underlying Freebase schema. In view of that, the detected communities are effectively used for the purpose of revealing unobserved or implicit domain relationships.**

Keywords—Linked Open Data; Community Detection; Freebase

I. INTRODUCTION

Freebase is a large, collaboratively database of cross-linked data developed by Metaweb Technologies [1]. Freebase has incorporated the contents of several large, openly accessible data sources, such as Wikipedia and Musicbrainz, allowing users to add data and build structure by adding metadata tags that categorise or connect items.

On the other hand, the massive amount of Freebase data raises an inevitable demand for effective data analysis and visualisation. Unlike other significant endeavours for exploring and visualising Freebase data such as "Thinkbase" [2] [3] and "GraphCharter" [14], the paper focused solely on Freebase schema. The paper adopted a graph-driven approach for representing the complex schema of Freebase. Furthermore, modularity-based analysis was utilised in order to detect communities in Freebase schema. The detected communities are used to explore the interrelations among Freebase domains. Specifically, we claim the following contributions:

- Utilising community detection in order to reveal unobserved or implicit domain interrelationships in Freebase schema, which has not been addressed before, to the authors' best knowledge.
- Exploring the densely connected domain communities in Freebase schema, based on the "Included Type" relationships.
- Identifying the highly interrelated domains of Freebase schema that tend to be located in numerous communities.
- Furthermore, the study provides methodological lessons concerning constructing Freebase schema as directed weighted graph using "Included Type" relationships and "Instance Count" property.

doi:10.15849/icit.2015.0082

II. METHODOLOGY

A. *Representation of Freebase Schema as Directed Weighted Graph*

The Freebase schema was constructed as a graph, where the graph is broken down into the following components:

1. Nodes: Each node in the schema graph represented a Freebase type. The total number of graph nodes reached 1,659.
2. Edges: Linking nodes with directed edges was realised by using the "Included-Type" relationships. For instance, since the "Author" type included the "Person" type, therefore a directed edge was constructed denoting "Author" as the source node, and "Person" as the destination node. The total number of directed edges was 2,837. Figure (1) depicts an example of the included-type relationship.

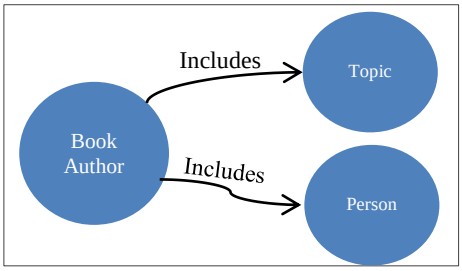

Fig. 1. An Example of how nodes were linked in the schema graph through directed edges that represent the included-type relationships.

3. Assignment of Edge Weights: The edge weight was used to indicate the relative influence of a source type on its included type. The "Instance Count", a schema property of Freebase, was considered for that purpose. Specifically, the edge weight is represented as the ratio of the source type instance count to the included type instance count. The edge weight is defined in equation (1) as follows:

$$W = (IC_{\text{Source Type}} / IC_{\text{Included Type}}) \quad (1)$$

Where

W → Edge Weight

$IC_{\text{Source Type}}$ → Instance Count of the Source Type (Source Node)

$IC_{\text{Included Type}}$ → Instance Count of the Included Type (Destination Node)

B. *Visualisation of the Schema Graph*

The constructed schema graph was utilised for the purpose of visualisation. Figure (2) illustrates the schema graph with emphasis on the highest degree nodes. The graph analysis and visualisations were conducted using Gephi [5].

Gephi is an open-source software for network exploration and manipulation. According to [14], Gephi modules can import, visualise, spatialise, filter, manipulate and export all types of networks. The visualization module uses a special 3D render engine to render graphs in real-time, using the computer graphic card. It can deal with large networks (i.e. over 20,000 nodes), because it was built on a multi-task model taking advantage of multi-core processors.

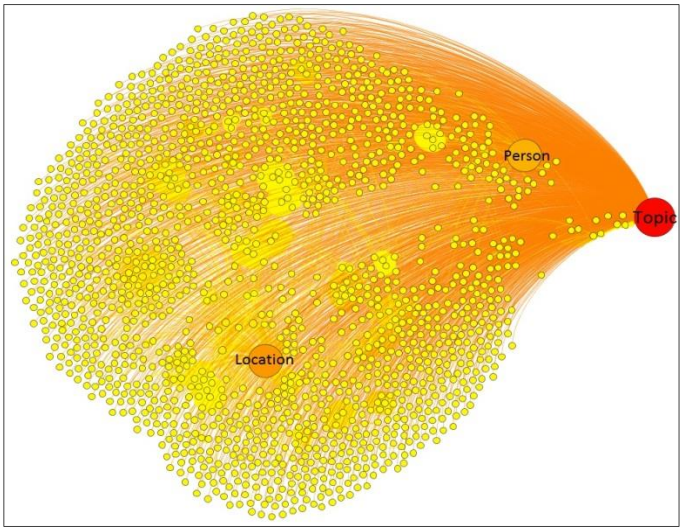

Fig. 2. Freebase schema graph with emphasis on significantly high ranked in-degree nodes. The rank of the node in-degree is represented as the node background colour ranging from yellow (lower in-degree) to red (higher in-degree). The edge directions are highlighted by the colour of source nodes.

C. *Minimisation of the Schema Graph*

The Schema graph was refined to present a higher view of the schema objects relationships, which is domain-based. The domain-based schema could provide a less complex graph providing an elevated perspective of Freebase schema objects interrelations. Moreover, the significantly lower number of Freebase domains (82) compared to that of Freebase types (1,659) directly contributed to decrease the complexity of the problem, and the following graph-based analysis.

For the purpose of schema minimisation, a new property needed to be added to Freebase schema, which is "Collective Instance Count". The collective instance counts were used to assign weights to edges of the minimised graph. Collective instance count accumulatively summed the instance counts of all types associated with a specific domain. For instance, the collective instance count of "Film" domain approximately reached 4,700,00 by adding up all the instance counts of the underlying types such as "Film director" , "Film actor" ," Film producer". The edge weight is defined in equation (2) as follows:

$$W = (CID_{\text{Source Domain}} / CID_{\text{Included Domain}}) \quad (2)$$

Where

W → Edge Weight

doi:10.15849/icit.2015.0082

CID $_{\text{Source Domain}}$ →Collective Instance Count of the Source Domain (Source Node)
CID $_{\text{Included Domain}}$→Collective Instance Count of the Included Domain (Destination Node)

As a result, the number of schema graph nodes decreased from 1,659 to 82. More importantly, the number of directed edges was reduced approximately by 90% from 2,837 to 274.

D. Visualisation of the Minimised Schema Graph

The minimised schema graph was re-visualised with respect to the domain-based perspective, as shown in figure (3). The graph nodes represent Freebase domains, and the directed edges represent the included-type relationships. In addition, figure (4) demonstrates the top 10 ranked Freebase domains by the in-degree measure.

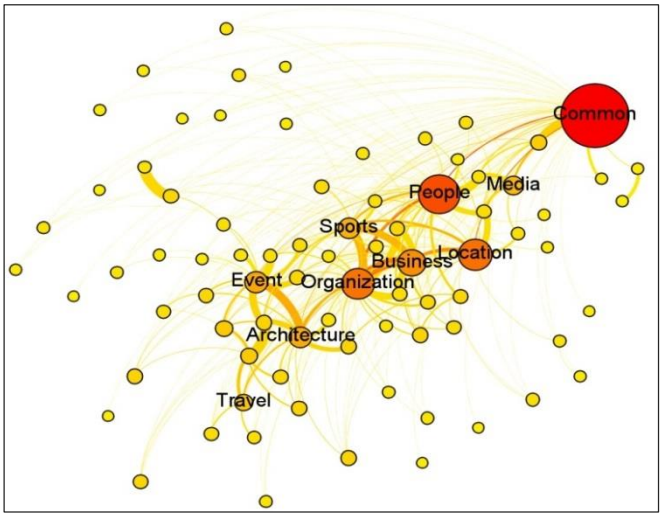

Fig. 3. Domain-based visualisation of Freebase schema graph, with emphasis on significantly high in-degree nodes. The rank of the node in-degree is represented as the node background colour ranging from yellow (lower in-degree) to red (higher in-degree). The edge directions are highlighted by the colour of source nodes.

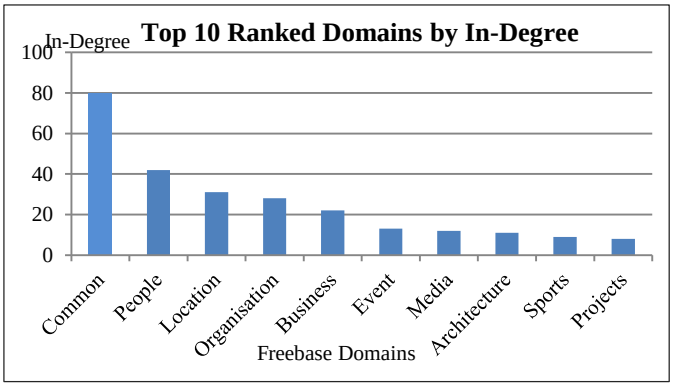

Fig. 4. Top 10 ranked Freebase domains by node in-degree. The "Common" domain has the significantly highest in-degree.

E. Normalising the Impact of High-Degree Nodes

The modularity-based analysis was adopted for detecting potential communities in the schema graph. However, the measure of modularity [6] is based on a principle that the connectivity within a community should be high, and the connectivity among communities should be low. Therefore, the negative impact of high-degree nodes should be normalised first before conducting the modularity analysis. The need for removing the higher degree nodes was acknowledged in a similar study [4] for summarizing large-scale database schemas using community detection as well.

Accordingly, the highest degree node was excluded from the schema graph, which represented the "Common" domain. As a result, the number of graph nodes and edges were reduced once again. The number of nodes decreased to 71, the exclusion of the "Common" domain resulted in the omission of other domains that had exclusive links to "Common". Eventually, the number of edges was reduced to 197.

F. Modularity-Based Analysis

The paper adopted the algorithm presented in study [7] for conducting the community detection, which was based on modularity measure. The selected algorithm was applied in different studies related to complex network analysis such as [10], [11], [12] and [13]. The modularity measure of weighted networks, which applies to the constructed Freebase schema graph, is defined in equation (3) according to [8]:

$$Q=\frac{1}{2m}\sum_{i,j}\left(A_{ij}-\frac{k_i k_j}{2m}\right)\delta(c_i,c_j) \qquad (3)$$

Where
A_{ij} →The weight of the edge between i and j
$k_i=\sum_j A_{ij}$ →The sum of edge weights attached to vertex i
c_i → The community to which vertex i is assigned
$\delta(u, v)$ is 1 if u = v and 0 otherwise, and $m=(½)\sum_{ij} A_{ij}$

The modularity-based analysis detected five densely connected communities. Figure (5) illustrates the five detected communities .Table (1) summarises the detected communities, the count of domains associated with each community and the included domains.

doi:10.15849/icit.2015.0082

TABLE.1 SUMMARY OF DETECTED COMMUNITIES.

Detected Community #	No. of Included Domains	Included Domains
1	26	Architecture, Travel, Amusement Parks, Zoos and Aquariums, Fashion; Clothing and Textiles, Military, American football, Olympics, Tennis, Skiing, Cricket, Event, Time, Aviation, Transportation, Spaceflight, Automotive, Projects, Theatre, Opera, Books, Law, Religion, Conferences and Conventions, Royalty and Nobility, Engineering
2	21	Education, Organization, Government, Language, Business, Digicams, Food & Drink, Soccer, Sports, Ice Hockey, Baseball, Basketball, Medicine, Computers, Meteorology, Biology, Astronomy, Location, Protected Places, Rail, Bicycles
3	17	Media, Film, Music, TV, Physical Geography, Visual Art, Video Games, Fictional Universes, Internet, Comics, Games, Awards, Hobbies and Interests, Geology, Periodicals, Comedy, People
4	2	Martial Arts, Boxing
5	2	Broadcast, Radio

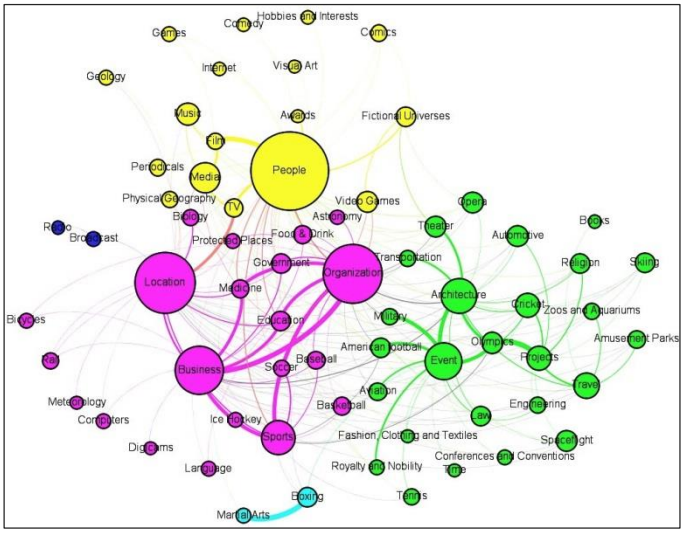

Fig. 5. Detected communities according to the modularity analysis. Each community is assigned a different colour for the purpose of demonstration.

G. Measuring Similarity Between Detected Communities and Freebase Categories

Freebase schema already includes a particular object as a grouping of related domains, which is "Category". The Freebase categories were considered as explicit communities to be compared with the implicit (detected) communities. However, the domains underlying Freebase categories could not be found explicitly neither on Freebase.com nor other reference, to the authors' best knowledge. Therefore, the domains of each category had to be extracted using MQL queries, below is an example of retrieving domains in "Science & Technology" category. Additionally, table (2) demonstrates the extracted domains of Freebase categories.

MQL Example: MQL query to retrieve domains of "Science & Technology" category:

```
[{
  "id": null,
  "name": null,
  "type": "/freebase/domain_profile",
  "category": {
    "id": "/en/science_technology" }
}]
```

TABLE. 2 FREEBASE CATEGORIES AND INCLUDED DOMAINS.

Freebase Category Name	Included Domains
Science & Technology	Medicine, Computers, Meteorology, Biology, Spaceflight, Internet, Astronomy, Chemistry, Geology, Engineering, Physics
Arts & Entertainment	Film, Music, Books, TV, Broadcast, Visual Art, Video Games, Theatre, Opera, Fictional Universes, Comics, Media, Games, Radio, Periodicals
Sports	Soccer, American football, Basketball, Sports, Ice Hockey, Baseball, Tennis, Cricket, Martial Arts, Olympics, Skiing, Boxing
Society	Education, Government, Language, People, Organization, Law, Religion, Awards, Conferences and Conventions, Influence, Library, Exhibitions, Celebrities, Royalty and Nobility
Products & Services	Food & Drink, Business, Digicams, Automotive
Transportation	Aviation, Transportation, Spaceflight, Boats,

doi:10.15849/icit.2015.0082

	Automotive, Rail, Bicycles
Time & Space	Location, Measurement Unit, Physical Geography, Time, Protected Places, Event
Special Interests	Architecture, Military, Travel, Amusement Parks, Zoos and Aquariums, Hobbies and Interests, Fashion-Clothing and Textiles, Symbols

Subsequently, the Jaccard similarity coefficient (Jaccard Index) was employed to measure the similarity between the implicitly detected communities and the explicitly defined categories by Freebase. The Jaccard index measures similarity between two finite sample sets as defined in equation (4) according to [9]:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (4)$$

III. RESULTS

The similarity measurement produced 40 Jaccard indices. Table (3) presents the values of Jaccard indices. In addition, figure (6) plots the Jaccard indices against the detected communities.

TABLE. 3 JACCARD SIMILARITY COEFFICIENTS OF DETECTED COMMUNITIES.

Community #	Jaccard Similarity Coefficients							
	Sc. & Tech.	Art & Ent.	Sport	Soc.	Prod, &Serv.	Trans.	Time & Space	Spec. Interest
1	**0.03**	**0.08**	**0.15**	**0.11**	**0**	**0.14**	**0.07**	**0.22**
2	**0.19**	**0**	**0.18**	**0.13**	**0.14**	**0.08**	**0.08**	**0**
3	**0.08**	**0.45**	**0**	**0.07**	**0**	**0**	**0.05**	**0.04**
4	**0**	**0**	**0.2**	**0**	**0**	**0**	**0**	**0**
5	**0**	**0.13**	**0**	**0**	**0**	**0**	**0**	**0**
	Average Similarity							
	0.06	**0.13**	**0.11**	**0.06**	**0.03**	**0.04**	**0.04**	**0.05**

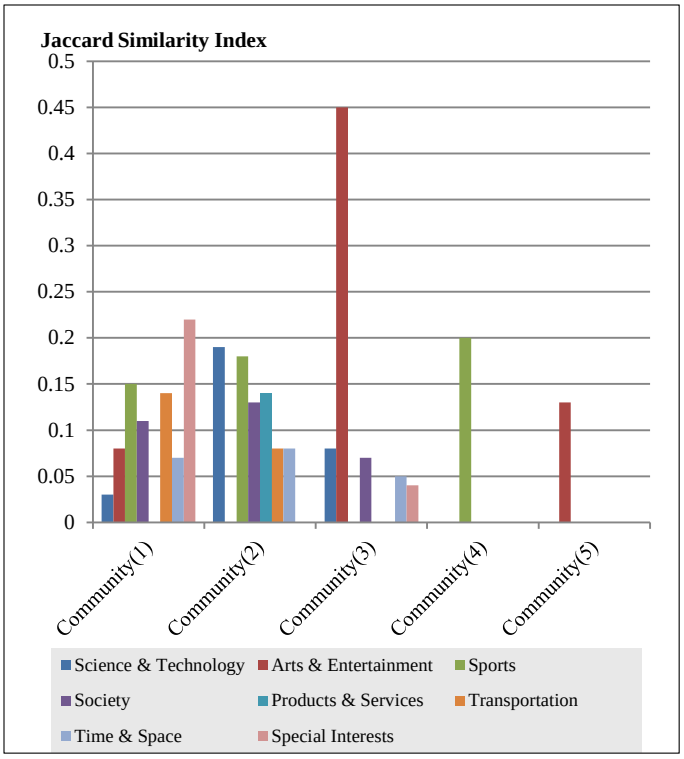

Fig. 6. Plotting Jaccard similarity indices against detected communities.

IV. DISCUSSION

Based on the Jaccard similarity measurements, the detected communities tended to have higher similarity with the categories of "Arts & Entertainment" and "Sports". However, identical or relatively large similarity was not expected, which can be justified that domain implicit interrelations are not explicitly established within Freebase categories. For instance, the first community included diverse domains from different Freebase categories, which are "Society", "Sports", "Time & Space", "Special Interests", "Transportation", "Arts & Entertainment", "Science & Technology". The diversity of domains included in the first community can depict the underlying interrelationships originating from the included-type relationships. However, the similarity indices can be considered as an indicator to the highly clustered communities, such as the third community.

Furthermore, the intensity of domain categories located in the detected communities could infer the interrelationships between Freebase domains. For example, the domains of "Sports" and "Society" categories can be considered to be highly involved or interrelated with other domains in diverse categories. On the contrary, "Products & Services" domains are exclusively located in an isolated community. Accordingly, the highly inter-linked domains are likely to be included in more communities. Figure (7) portrays the overlaps between the five detected communities.

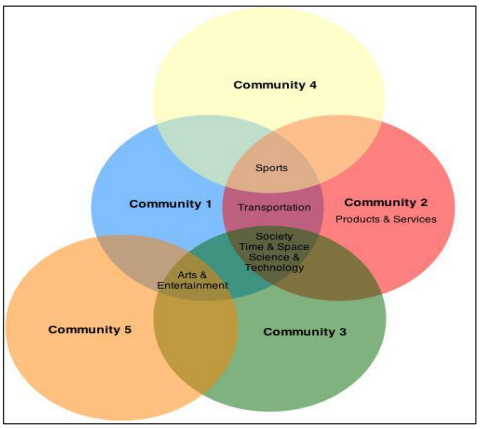

Fig.7. Overlaps between the detected communities. Highly-interrelated domains are located in intensively intersected areas, while less inter-related domains are located in fewer communities.

doi:10.15849/icit.2015.0082

V. LIMITATIONS OF THE METHODOLOGY

The adopted methodology depended mainly on two particular properties of Freebase schema for constructing the schema graph, which are "Included Types" and "Instance Count". Therefore, it might not be possible to generalise that methodology, to build other schema graphs, unless similar schema properties are available. However, the methodology can still be useful with Freebase case for the purpose of graph-based analysis or visualisation.

VI. CONCLUSIONS

In the first instance, the paper presents a graph-driven approach to analyse and visualise the large-scale schema of Freebase. The Freebase schema is represented as a directed weighted graph. Initially, the schema graph is constructed using Freebase types, included-type relationships and instance count property. Afterwards, the schema graph is minimised and restructured with respect to Freebase domains. Eventually, the impact of high-degree nodes has been normalised by excluding those nodes from the schema graph.

Secondly, modularity-based analysis is utilised to detect potential communities in Freebase schema graph.The modularity analysis could identify five densely connected communities. The Jaccard similarity indices are used to measure the similarity between the implicitly detected communities and the explicitly defined categories by Freebase. The similarity measurements can indicate that "Arts & Entertainment" and "Sports" categories have higher similarity with the detected communities. Furthermore, the overlaps between the detected communities can detect the highly inter-linked domains in Freebase schema, such as the domains of "Society" category. Hence, the community detection is demonstrated as an effective method that can reveal unobserved or implicit relationships within complex graph-based schemas, such as Freebase.

REFERENCES

[1] Arrison, Thomas, and Scott Weidman, eds. Steps Toward Large-Scale Data Integration in the Sciences:: Summary of a Workshop. National Academies Press, 2010.

[2] Hirsch, Christian, John C. Grundy, and John G. Hosking. "Thinkbase: A Visual Semantic Wiki." In International Semantic Web Conference (Posters & Demos). 2008.

[3] Hirsch, Christian, John Hosking, and John Grundy. "Interactive visualization tools for exploring the semantic graph of large knowledge spaces." In Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW2009), vol. 443. 2009.

[4] Wang, Xue, Xuan Zhou, and Shan Wang. "Summarizing large-scale database schema using community detection." Journal of Computer Science and Technology 27, no. 3 (2012): 515-526.

[5] https://gephi.github.io/

[6] Newman, Mark EJ, and Michelle Girvan. "Finding and evaluating community structure in networks." Physical review E 69, no. 2 (2004): 026113.

[7] Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. "Fast unfolding of communities in large networks." Journal of Statistical Mechanics: Theory and Experiment 2008, no. 10 (2008): P10008.

[8] Newman, Mark EJ. "Analysis of weighted networks." Physical Review E 70, no. 5 (2004): 056131.

[9] Cesare, Silvio, and Yang Xiang. Software similarity and classification. Springer Science & Business Media, 2012.

[10] Lancichinetti, Andrea, and Santo Fortunato. "Community detection algorithms: a comparative analysis." Physical review E 80, no. 5 (2009): 056117.

[11] Porter, Mason A., Jukka-Pekka Onnela, and Peter J. Mucha. "Communities in networks." Notices of the AMS 56, no. 9 (2009): 1082-1097. Fortunato, Santo, "Community detection in graphs." Physics Reports 486, no. 3 (2010): 75-174.

[12] Rubinov, Mikail, and Olaf Sporns. "Complex network measures of brain connectivity: uses and interpretations." Neuroimage 52, no. 3 (2010): 1059-1069.

[13] Tu, Ying, and Han-Wei Shen. "GraphCharter: Combining browsing with query to explore large semantic graphs." In Visualization Symposium (PacificVis), 2013 IEEE Pacific, pp. 49-56. IEEE, 2013.

[14] Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. "Gephi: an open source software for exploring and manipulating networks." ICWSM 8 (2009): 361-362.