Sociotechnical Behavior Mining: From Data to Decisions? Papers from the 2015 AAAI Spring Symposium

An Online Platform for Focal Structures Analysis

Analyzing Smaller and More Pertinent Groups Using a Web Tool

Fatih Şen, Naga Nagisetty, Teeradache Viangteeravat
fsen@uthsc.edu, nnagiset@uthsc.edu, tviangte@uthsc.edu
Children's Foundation Research Institute
Le Bonheur Children's Hospital
Memphis-TN, USA

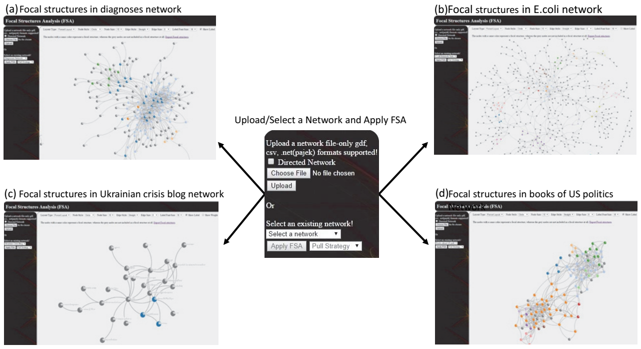
Nitin Agarwal
nxagarwal@ualr.edu
Department of Information Science
University of Arkansas at Little Rock
Little Rock-AR, USA

Abstract

Biological networks are increasingly becoming available for the researchers and practitioners to mine and analyze meaningful structures. The traditional approaches, such as detecting communities/clusters, do not focus on smaller and more relevant groups of individuals. We design and develop a powerful web tool (http://focalstructures.net) to identify focal structures, influential sets of individuals, in a given network. This paper demonstrates the online usage of the Focal Structures Analysis (FSA) approach and aims to give a brief understanding what a focal structure is about. Besides biological networks, the FSA approach can be experimented on different application domains (such as social networks) as well. The web tool helps researchers and practitioners to mine smaller and meaningful structures in an easier and more accessible way.

Introduction

Analyzing biological networks by extracting biologically meaningful structures can help with understanding of the function of the organism. Studying such complex networks plays an important role to understand how their interactions function within a living cell. Many computational algorithms, such as authoritative approaches or community identification methods, have been proposed to detect and analyze the structure and functional units of such networks in which vertices represent individuals (genes, proteins, etc.) and edges represent the interactions among them (Girvan and Newman 2002; Barabási and Oltvai 2004; Adamcsek et al. 2006; Mete et al. 2008). Those methods help researchers to characterize and understand higher level functions of the living organisms. Yim et al develops tYNA (Yip et al. 2006), a web system for managing, comparing, and mining multiple networks. Another major study is Cytoscape (Shannon et al. 2003), which is an open-source software platform for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other functional genomics data. However, none of them focuses on in smaller and more pertinent groups in networks. To identify and analyze such structures, the FSA (Focal Structures Analysis) approach has been developed in our previous studies (Şen et al. 2013; 2014). We design and implement a web tool (http://focalstructures.net) to demonstrate the usage of the FSA approach. The online system (still under development) aims to give an understanding to researchers how to analyze focal structures in a given network in an easier and more accessible way.


(a)Focal structures in diagnoses network
(b)Focal structures in E.coli network
Upload/Select a Network and Apply FSA
(c) Focal structures in Ukrainian crisis blog network
(d)Focal structures in books of US politics
Focal Structures Analysis (FSA)
Directed Network
Choose File
No file chosen
Upload
Or
Select a network
Apply FSA
Pull Strategy

Upload a network file-only gdf, csv, .net(pajek) formats supported!

Select an existing network!

Figure 1: Four different networks with identified focal structures. After uploading a network or selecting an existing network in the system, focal structures are identified by pressing "Apply FSA" button. (a), (b), (c), (d) are showing focal structures in diagnoses network, E.coli, Ukrainian Crisis Blog and books about US politics networks, respectively.

Figure 1 shows an example of focal structures in various networks. The user loads a network and clicks the "Apply FSA" button to identify focal structures. For example, Escherichia coli (E.coli) intermediary metabolism is a metabolic network (Figure 1b) which contains 513 vertices representing compounds and 750 edges representing the interactions among them. The vertices with a same color represent a focal structure, whereas the grey nodes are not included in a focal structure at all.

The FSA web tool shares some elements with some other

Table 1: Comparison between FSA web tool and other existing systems

	Cytoscape	Gephi	tYNA	N-Browse	FSA Web Tool
Main purpose	Visualization	Visualization	Analysis	Visualization	Analysis
System	Desktop	Desktop	Web	Web	Web
Stat. Cal.	Plug-in	Built-in	Built-in	No	No (Future Work)
Plug-in Support	Yes	Yes	No	No	No (Future Work)
Focal Structures Finding	No	No	No	No	Yes

network analysis and visualization systems, such as Cytoscape (Shannon et al., 2003), Gephi [1], tYNA [2], JUNG [3], and N-Browse [4], but also offers some additional features such as the analysis of focal structures. A comparison between the FSA web tool and other existing systems is given in Table 1.

The Architecture of the Online System

The architecture of the web site for analyzing focal structures consists of web user-interface and server-side. The Java Servlet Pages (.jsp) technology is utilized on the web interface, whereas Java programming language is used on the server-side for computing the FSA method. The focal structures are obtained on the server side and given to the browser as a json format. The user web interface receives that json-formatted dataset from the server to visualize it using d3.js [5] library.

The Methodology

A focal structure is an influential group of individuals in a network. Those individuals do not have to be strongly connected and may not be influential enough on their own but by acting together they form a compelling power. Depending on the size of a network, a focal structure may be a community, but it is most likely different, because focal structures are smaller and more relevant groups of individuals. Communities are mostly larger networks and focal structures represent networks interacting in smaller and more pertinent groups.

To identify focal structures, we develop an algorithm (Algorithm **??**), which utilizes a clustering coefficient to decide which vertices from the neighborhood are to be included in a focal structure at each step. First, we compute the clustering coefficient values of each vertex and the mean of the computed values in the graph. Then, we utilize them for pair-wise comparisons. For each pair-wise comparison, if the clustering coefficient of both vertices is less or greater than the mean, then those vertices are included in the same structure. This process leads to generating structures of vertices with closer clustering coefficient values.

$$C_n = \frac{2e_n}{k_n(k_n-1)} \qquad (1)$$

where k_n is the number of neighbors of n and e_n is the number of connected pairs between all neighbors of n. This variant is defined by Duncan J. Watts and Steven Strogatz (Watts and Strogatz 1998) as the local clustering coefficient.

The focal structure id (index) is obtained before any pair-wise comparison. It is used to add the neighbor vertex v_j to that focal structure which includes the vertex v_i. Please note that if a vertex does not have more than one neighbor, it is not considered to be included in a structure even though its clustering coefficient might be 1.

Diagnoses Network and Use Case Study

Comorbidities are proved to be associated with worst outcomes, increase in complexity of and cost of treatment [6]. Defining comorbidities is an important factor in deciding the treatment pathway, often is complex and diagnoses specific [7]. Secondary use of electronic medical record data is a trending methodology of using historical health records of similar patient profiles to understand conditions, progression and outcomes.

In an attempt to perform a data driven approach to find most common co-occurring diagnoses or comorbidities, we extract all ICD-9 codes for secondary diagnoses of patients from our Electronic Medical Records (EMRs) - derived Pediatric Research Database (PRD) (Viangteeravat and Nagisetty 2014; Nagisetty et al. 2014). PRD is a de-identified pediatric health information designed to provide researchers and physicians with a clinically rich and useful data source, such as cohort discovery and visual analytics. We build an undirected network of co-occurring pairs of diagnoses with at least a hundred occurrences as a first attempt, where each vertex represents a diagnoses and edge represents the patient(s) carrying those diagnoses. This network of diagnoses consists of 206 vertices and 600 edges. As shown in Figure 2, we select the diagnoses network and apply FSA to the network to identify focal structures. These attempts lead to identification of influential sets of diagnoses which are commonly occurring in similar patient profiles.

Some examples of identified focal structures are shown below.

1) Focal structure id-5: “Neurogenic bladder NOS:596.54”, “With hydrocephalus, spina bifida, unspecified region:741.00”, “Spina bifida, without mention of hydrocephalus, unspecified region:741.90”

In this focal structure, all the diagnoses are related to neural and spine development and the structure uniquely identify issues with bowel and bladder movements.

2) Focal structure id-13: “Attention deficit disorder of childhood with hyperactivity:314.01”, “Infantile autism, current or active state:299.00”

[1] http://gephi.github.io/

[2] http://tyna.gersteinlab.org/tyna/

[3] http://jung.sourceforge.net/

[4] http://nematoda.bio.nyu.edu:8080/NBrowse/N-Browse.jsp?last=false

[5] http://d3js.org/

[6] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2713155/

[7] http://www.sciencedirect.com/science/article/pii/S1059131110001123

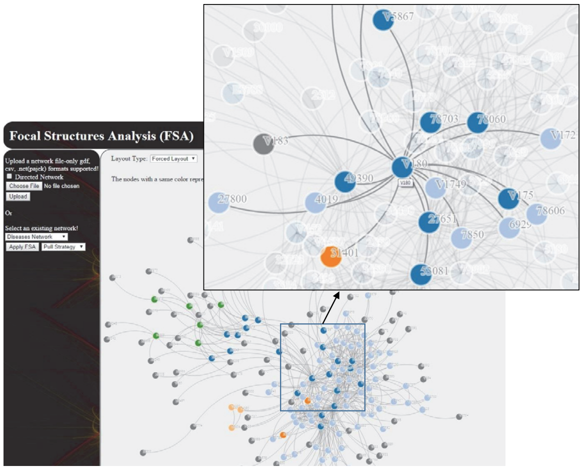
In this focal structure, all the diagnoses are related to attention disorder and autism, which is a very common in the autictic population.

```
Data: G = (V,E), where V is the set of vertices and E is the set of edges.
Result: fList: A list of focal structures

function GETFOCALSTRUCTURES(G)
    $f$: A focal structure object of vertices and add-remove functions
    Generate fList
    Compute the clustering coefficient of each vertex in the graph G.
    Compute the mean $c_m$ of the clustering coefficient values
    i ← 1
    while i ≤ size of vertices do
        Obtain the vertex $v_i$
        Obtain the clustering coefficient of $c_i$ of $v_i$
        if vertex $v_i$ does not belong to any structure then
            Obtain a new $f_i$ which is a focal structure of vertices
            Add vertex $v_i$ to the $f_i$
            Obtain the index of $f_i$ as i
            Add $f_i$ to the ith position of fList
        end
        else
            Obtain the index (focal structure id) that vertex $v_i$ belongs to
        end
        j ← i + 1
        while j ≤ size of vertices do
            Obtain the vertex $v_j$
            if $v_i$ and $v_j$ are neighbors and $v_i$ has more than one neighbor then
                Obtain the clustering coefficient $c_j$ of $v_j$
                if ($c_i < c_m$ and $c_j < c_m$) or ($c_i > c_m$ and $c_j > c_m$) then
                    Add the vertex (neighbor) $v_j$ to the focal structure $f_{index}$ in which the vertex $v_i$ is involved.
                end
            end
            increment j by 1
        end
        increment i by 1
    end
    return fList
end function
```

Algorithm 1: Obtaining focal structures based on clustering coefficient values

3) Focal structure id-17: “Diagnosis - External cause status NOS:E0009”, “Contusion of face, scalp, and neck except eye(s):920”, “Unspecified activity: E030”, “Burn (any degree) involving less than 10 percent of body surface with third degree burn of less than 10 percent or unspecified amount:948.00”, “Diagnosis - Tackle football (Activities involving American tackle football):E0070”, “Diagnosis - Sports acc w/o sub fall (Striking against or struck accidentally by objects or persons in sports):E9170”.



Focal Structures Analysis (FSA)

Upload a network file-only gdf, csv, .net(pajek) formats supported!

☐ Directed Network

Choose File No file chosen

Upload

Or

Select an existing network!

Diseases Network ▾

Apply FSA | Pull Strategy ▾

Layout Type: Forced Layout ▾

The nodes with a same color repre

Figure 2: A use case study with diagnoses network.

In this focal structure, all the diagnoses are sport related injuries or activities and burns that have relation with contusion of face and scalp.

Discussion and Future Work

This paper demonstrates a powerful web tool of mining and analyzing influential sets of individuals, i.e. focal structures. Since FSA is evaluated in previous studies, here we focus more on how to benefit from it in an online system. Please note that FSA is not an algorithm, it is more an approach for which many different algorithms can be developed. However, it is not restricted only for identifying focal structures, it also has the potential of the identification and analysis of motifs, communities/clusters and authoritative individuals as a future improvement. Therefore, we also take into account to answer the big picture and question “How can we develop an easy-use, powerful and more accessible tool so that researchers or practitioners can use it for their studies without having much technical knowledge?”.

When FSA applied to the existing “Protein Complexes” biological network, it cannot be visualized on the web due to its large size (43,540 edges-interactions among 4,715 vertices-proteins). The cutting-edge d3.js technology is unable to visualize that many interactions on the web. However, focal structures can still be obtained on the back-end.

The discussion of the web tool presented here is based more on biological network analysis, however, it can serve different types of application domains as well. For example, it can contribute to social network analysis which is an emergent trend for researchers in recent years. As an example of an analysis of a social network, the current version of the web tool includes a blog network of the 2014 Ukrainian crisis (Figure 1c) which began in November 2013, when President Viktor Yanukovych rejected a deal for greater integration with the European Union. Graham W. Phillips, a 35-year-old British journalist and a blogger covered the 2014 Ukraine crisis, became a growing star on Kremlin-owned media. After applying FSA to this dataset, the online system identifies only one focal structure in which he is involved as well.

Currently, the system provides researchers four main functionalities:

- Supporting biological and social networks with undirected networks with .gdf (GUESS) network file format.
- User-friendly design and easy use mechanism.
- Efficient interactive graphical user interface.
- Analyzing focal structures.

Recall that the web tool is still under development. As future work, we plan to improve our system with the following features:

- Multi-network and file type support: Supporting both directed and undirected networks with various network file formats (.net-pajek, .gexf, .csv, etc.).
- Exporting the output: Exporting and saving the experiment results.
- Statistics Calculation: Providing statistics of the network properties.
- Multi-structure support: Identifying and analyzing motifs, communities/clusters, and influential individuals.
- Application Programming Interface (API) support: Providing third party developers/researchers to utilize the FSA approach for their applications.

After completing the web tool by implementing all the features mentioned above, we plan to make the source codes of the web tool publicly available so that researchers can download the full package of the whole online system and run it on their local environment. We believe that this is extremely important for the biomedical and bioinformatics researchers, because the fact of getting an approval from the institutional review board (IRB), which protects the rights and welfare of humans participating as subjects in a research study, makes it harder for them to perform analysis on servers located outside their institution. For example, as researchers at Children's Foundation Research Institute (CFRI), we could not use http://focalstructures.net, because the hosting server is located outside our institution's firewall and the diagnoses dataset, collected from our PRD system, requires the IRB approval. If placing the data in the http://focalstructures.net environment comes to necessity, we will work to modify the IRB accordingly and will submit the necessary data sharing agreements for approval prior to sharing any data. Instead, we run it on our local environment (http://localhost:8080/fs/) for the analysis. Providing researchers such an option would save their time and allow them to perform safer analysis as well.

Acknowledgment

The authors thank the Children's Foundation Research Institute of Le Bonheur Children's Hospital and University of Arkansas at Little Rock for the support and collaboration on this work. This research is funded in part by the U.S. Office of Naval Research under award number N000141410489.

Conflict of Interest

The authors declare no conflict of interest.

References

Adamcsek, B.; Palla, G.; Farkas, I. J.; Derényi, I.; and Vicsek, T. 2006. Cfinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 22:1021–1023.

Barabási, A.-L., and Oltvai, Z. N. 2004. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* 5(2):101–113.

Şen, F.; Wigand, R. T.; Agarwal, N.; Mahata, D.; and Bisgin, H. 2013. Identifying focal patterns in social networks. 105–108. Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference.

Şen, F.; Wigand, R. T.; Agarwal, N.; Mete, M.; and Kasprzyk, R. 2014. Focal structure analysis in large biological networks. volume 70. IACSIT Press.

Girvan, M., and Newman, M. E. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99(12):7821–7826.

Mete, M.; Tang, F.; Xu, X.; and Yuruk, N. 2008. A structural approach for finding functional modules from large biological networks. *BMC Bioinformatics* 9(Suppl 9):S19.

Nagisetty, N.; Huang, E. Y.; Wade, G.; and Viangteeravat, T. 2014. Building a knowledge base to assist clinical decision-making using the pediatric research database (prd) and machine learning: a case study on pediatric asthma patients. *BMC Bioinformatics* 15(Suppl 10):17.

Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; and Ideker, T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 13(11):2498–2504.

Viangteeravat, T., and Nagisetty, N. S. V. R. 2014. Giving raw data a chance to talk: A demonstration of exploratory visual analytics with a pediatric research database using microsoft live labs pivot to promote cohort discovery, research, and quality assessment. *Perspectives in Health Information Management* 11(Winter).

Watts, D. J., and Strogatz, S. H. 1998. Collective dynamics of small-world networks. *Nature* 393(6684):440–442.

Yip, K. Y.; Yu, H.; Kim, P. M.; Schultz, M.; and Gerstein, M. 2006. The tyna platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. *Bioinformatics* 22(23):2968–2970.