bioRxiv preprint doi: https://doi.org/10.1101/168757; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CINNA: Deciphering Central Informative Nodes in Network Analysis

Minoo Ashtiani[1], Mohieddin Jafari[1]

[1]Drug Design and Bioinformatics Unit, Medical Biotechnology Department, Biotechnology Research Center, Pasteur Institute of Iran, 69 Pasteur St, PO Box 13164, Tehran, Iran

Abstract

Motivation: With the advancement of data mining technology, created networks from various contexts become more complex and larger, which makes them more difficult to figure out. Recognizing nodes that can influence on the whole network, helps us to comprehend networks easier and faster and so facilitates the process of network analysis. Since several criteria based on the network topological features are defined for identifying influential nodes, we need to know which measure rightfully points at the central nodes in special network.

Results: The CINNA R package conveniently has brought together all required methods for network centrality analysis. It contains network component segregation, calculation and prioritizing centralities, along with clustering and visualization functions.

Availability: *CINNA package* is freely available from the R project at http://cran.r-project.org/, http://jafarilab-pasteur.com/content/software/CINNA.html.

Contact: m_jafari@pasteur.ac.ir

1 BACKGROUND

Nowadays, the importance of finding basic and influential nodes of diverse networks is obviously determined. As we go forward, networks in the context of different sciences come to be more enormous and complex. Therefore, meticulously analyzing each node of a network is costly and time-consuming and somehow impossible. So, various methods has been found to help us detect the essential nodes. Centrality measure is the procedure of figuring out the influential nodes through a network. Several centrality measures have been known to detect most influential nodes using different approaches [1]. Most of these measures can be characterized based on network features including direction, bipartition or weight.

But the primer question is that, which one of the centrality measures can precisely point at essential nodes of networks. Many studies have been conducted to compare the behavior of centralities within several networks during these years. Koschützki and Schreiber came to an independent behavior among centrality measures according to various visualization comparison methods [2]. At the same time, Borgatti and Everett found that network topology can strongly affect the accuracy of centrality measures values [3]. Furthermore, Cong Li *et al*. tried to understand the dependency of several centrality indices based on correlation analysis [4]. Beside these results, recently, we depicted diverse centrality indices have different levels of information about central nodes. More precisely, comparing 27 centrality measures in 14 distinct networks showed that properly detecting central nodes requires utilizing dimensional reduction methods such as principal components analysis (PCA), [5].

bioRxiv preprint doi: https://doi.org/10.1101/168757; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To best of our knowledge, no specified pipeline has been introduced for prioritizing centrality indices within a given network. The CINNA R package is developed to apply network centrality analysis and designate more appropriate measure. It is aimed to be applicable on every network models within various scientific areas. In other word, CINNA is a collection of functions which can be used to compute, compare, assort and visualize network centrality measures together.

2 FUNCTIONALITY

Centrality measure values depend on the topological network features. Therefore, having a general overview of network topology and measurable indices would be a great help to figure out the real central nodes in network analysis [5].

Most of centrality measures require strongly connected networks [6].Then, functions for segregating components of a network are available for various formats such as igraph, network, adjacency matrix or edge list. It is also provided special functions to figure out bipartite networks and apply centrality analysis to one of the projections.

CINNA package provides all appropriate centralities based on the network structure, which can be characterized to undirected-unweighted, undirected-weighted, directed-unweighted and directed-weighted graph. Moreover, using this package, researchers are able to limit the centrality measure computation based on their preferences.

In order to distinguish appropriate and most informative centralities, CINNA has prepared PCA [7] and t-Distributed Stochastic Neighbor Embedding (t-sne) algorithm [8]. Both of them are dimensionality reduction approaches respectively in terms of linear and non-linear analysis. In PCA method, using contribution criteria, it can be determined which centrality index contains more information and so, which one can describe central nodes of the network more accurately [5]. On the other hand, t-sne algorithm supplies cost criteria for acquiring most informative centralities. In these methods, centralities which have the highest information about the central nodes are assorted and can be visualized. Hence, user can be able to compare computed centrality results in terms of contribution or cost score.

Also, different visualization techniques for comparing results of centralities are provided including heat-map, dendrogram and pairwise scatter plot. The whole network can be illustrated by specifying a centrality whereas the sizes of vertices indicate the centrality values.

The unified manual can be accessed using the R command `help (package=CINNA)`. In addition, an immense step-by-step tutorial of all substantial features of the package can be approached using the command `browseVignette ("CINNA")`.

3 Exemplary uses of CINNA

As an example, we used Zachary network [9] to briefly glance on CINNA functionality. It can be loaded via

```
> data (zachary)
```

bioRxiv preprint doi: https://doi.org/10.1101/168757; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and is now kept in a variable called `zachary`. The giant component of the network [10] can be extracted using

```
> giant.component.extract (zachary)
```

which is an igraph class [11]. In order to declare proper centrality measures depending on the network structure,

```
> prop.cent <- proper.centralities (zachary)
```

which returns a vector including all proper centralities (44 indices) that are applicable on the Zachary network.

To compute a set of centralities mentioned in `prop.cent` object, the below function can be used,

```
> calc.cent <- calculate.centralities (zachary, except=
prop.cent[4:41])
```

Centralities with highest information level can be acquired for instance according to PCA data reduction,

```
> pca.centralities (calc.cent)
```

The corresponding results are illustrated on Figure 1. As it shown "Bonacich power centralities of positions" [12] has the highest contribution level among six calculated centralities For accessing the Eigen and contribution values calculated in PCA, the below function can be useful:

```
> summary.pca.centralities (calc.cent)
```

The CINNA package currently includes about 23 user-level functions along with 5 natural network examples [9, 13-16], which helps the user to have a different experience from the centrality network analysis. An essential part of the advancement process is user feedback. We appreciate receiving all suggestions and comments from the users.

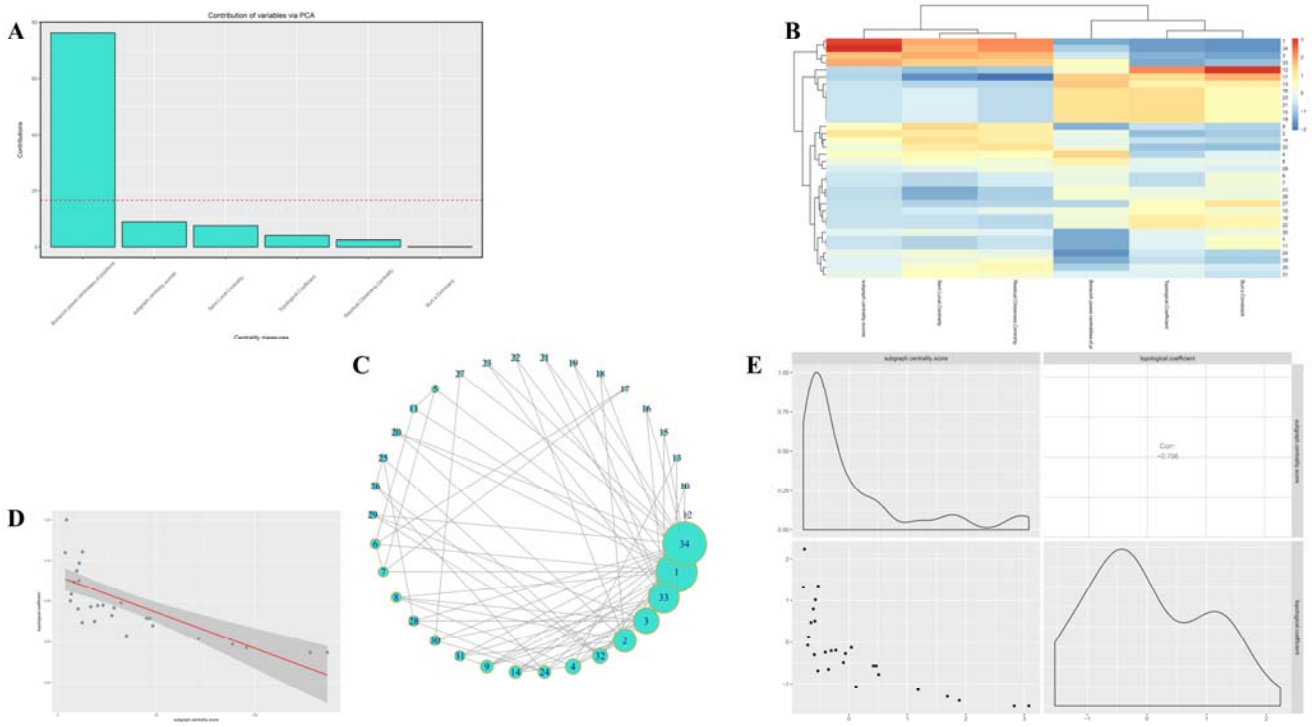

Figure 1: Exemplary uses of CINNA on Zachary network. A) PCA contribution bar plot. Bonacich power centralities of positions has the highest level of contribution and it is only centrality that come to up of the threshold which has been shown as the red line. B) Heat map of nodes centrality measure values. Colors spectra from blue to red represent nodes that have lowest to highest centrality values. C) Zachary graph visualization. Size of nodes correspond to the value of degree centrality. D) Association plot between two centrality measures. Linear relationship between "topological coefficient" and "subgraph centrality score" has been shown using a red line. E) Correlation plot among two centrality measures. This plot has four separated parts. Top left and top bottom right belong to respectively "subgraph centrality score" and "topological coefficient". Top right part visualize the correlation among the variables based on Pearson coefficient. At last, bottom left illustrates scatterplot of variables.

bioRxiv preprint doi: https://doi.org/10.1101/168757; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/168757; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements

The authors wish to thank Reza Karbalaei for designing the logo of CINNA package, which can be seen on http://www.jafarilab-pasteur.com/content/software/CINNA.html . Also, we acknowledge Ali Salehzadeh-Yazdi and Mahdi Jalili for suggestions and feedback.

References

1. Jalili, M., et al., *CentiServer: a comprehensive resource, web-based application and R package for centrality analysis.* PloS one, 2015. **10**(11): p. e0143111.
2. Junker, B.H., D. Koschützki, and F. Schreiber, *Exploration of biological network centralities with CentiBiN.* BMC bioinformatics, 2006. **7**(1): p. 219.
3. Borgatti, S.P. and M.G. Everett, *A graph-theoretic perspective on centrality.* Social networks, 2006. **28**(4): p. 466-484.
4. Li, C., et al., *Correlation between centrality metrics and their application to the opinion model.* The European Physical Journal B, 2015. **88**(3): p. 65.
5. Ashtiani, M., et al., *Revisiting the General Concept of Network Centralities: A Propose for Centrality Analysis in Network Science.* bioRxiv, 2017: p. 149492.
6. De Nooy, W., A. Mrvar, and V. Batagelj, *Exploratory social network analysis with Pajek*. Vol. 27. 2011: Cambridge University Press.
7. Abdi, H. and L.J. Williams, *Principal component analysis.* Wiley interdisciplinary reviews: computational statistics, 2010. **2**(4): p. 433-459.
8. Maaten, L.v.d. and G. Hinton, *Visualizing data using t-SNE.* Journal of Machine Learning Research, 2008. **9**(Nov): p. 2579-2605.
9. Zachary, W.W., *An information flow model for conflict and fission in small groups.* Journal of anthropological research, 1977. **33**(4): p. 452-473.
10. Newman, M., *Networks: an introduction*. 2010: Oxford university press.
11. Csardi, G. and T. Nepusz, *The igraph software package for complex network research.* InterJournal, Complex Systems, 2006. **1695**(5): p. 1-9.
12. Bonacich, P., *Power and centrality: A family of measures.* American journal of sociology, 1987. **92**(5): p. 1170-1182.
13. Barneh, F., M. Jafari, and M. Mirzaie, *Updates on drug-target network; facilitating polypharmacology and data integration by growth of DrugBank database.* Brief Bioinform, 2016. **17**(6): p. 1070-1080.
14. Felleman, D.J. and D.C. Van Essen, *Distributed hierarchical processing in the primate cerebral cortex.* Cerebral cortex (New York, NY: 1991), 1991. **1**(1): p. 1-47.
15. KONECT, *Rhesus Network Dataset*. 2016.
16. KONECT, *Kangaroo Network Dataset*. 2016.