The Problem of Ego Networks in Humanities Data

Melanie Conroy[1]

[1]University of Memphis

1 Introduction

Ego networks present a problem for humanities network research that aims to say something broader about the society or group of which the individual is a part. Many smaller humanities network datasets are built around one or two figures who are not representative of the larger historical networks to which they belong. Collections of letters, for example, often include only the correspondence of one person or a small group, even when the collections are quite large (Comsa et al., 2016; Edelstein et al., 2017). Ego networks are often called "personal networks" or "egocentric networks" in the wider network analysis literature. Within humanities network analysis, the problem of the egocentricity of networks has perhaps been less thematized because the focus on people and their connections - such as writers, artists, or historical figures - is embedded in many humanities fields as much as in digital humanities. I would like to propose, however, that reliance on "personal" or "ego" network data is a problem for humanities datasets and that this problem is structural; thus, it does not disappear when datasets are compiled or combined into larger data resources.

In this presentation, I explore ways to make visualizations of ego networks and other unrepresentative networks more useful for network analysis. While noting that we cannot overcome the problem of data bias, nor eliminate the outsized influence of ego networks in humanities network datasets, we can be more aware of the ways that data are biased. One way to be more careful about the reliance of humanities datasets on ego networks is to be examine the structural characteristics of ego networks that appear in both analytics and visualizations.

2 Humanities Data and Ego Networks

The ubiquity of ego networks is not merely a problem because the individual items in the collection are not representative of a broader set of objects or people, but also because the relational data are not representative of the connections within that larger group. Thus, the visualization of networks that center around one individual or small group will often serve only to emphasize the centrality of that group or individual, especially when researchers are using out-of-the-box tools or visualizing networks without reflecting on the underlying data.

When archives are curated around or by individuals, networks derived from them will be more likely to over-represent the people at the heart of the archive - whether the subjects of the archive, the collectors of the archival documents, or even the curators themselves (Conroy and Elo, 2020). In very large collections, biases are often disguised by the large number of ego networks in the dataset, despite the fact that they are just as present in collections of ego networks.

When we recognize collections of ego networks in data, there are many forms of valid analysis that open up. For one thing, we can compare the structure of ego networks - for example, by using small multiples to visually compare even large numbers of ego networks. For another, we can measure similarity between smaller networks, such as co-occurrence networks, mathematically to find ego networks with higher and lower similarity to one another (Wang et al., 2020).

3 Visualizing Ego Networks

Due to the distortions arising from how data are gathered, large collections of ego networks cannot tell us very much about the structure of larger groups like societies, or even large social networks. We can use networks and network graphs to explore societies, kinds of connections, or possible connections, without those connections being statistically representative of some underlying structure.

The characteristics and structures of ego networks are relatively well understood within the network analysis community (Freeman, 1982; Gupta et al. 2015). We can use these characteristics as a baseline for the discussion of patterns that we find within humanities datasets. The characteristics of ego networks that I think are most relevant to the analysis of smaller curated datasets are as follows:

- personal network density
- the number of communities
- connectors that link the ego network to others
- the extent to which individual ego networks are connected to other ego networks

The structure of ego networks, as focalized on one individual, means that they are not very good at assessing the importance, or the centrality, of that individual. If we collect data about prolific letter writers or socialites, we will find to no surprise that they appear as highly central to the network in which they occupy the center, and that they therefore have a high degree (meaning many connections to others within this network). In my own work on French salons of the eighteenth and nineteenth century, the hosts of these meetings tend to have the highest degree, since the measured connections within the salon network are the documented attendance at the salons. For example, one of the highest degree nodes in my 19th-Century Networks database is the very well connected French socialite and comtesse Élisabeth Greffulhe, an artefact of the data collection focused on important literary salons in Paris.

Figure 1 shows the ego network of Mme Greffulhe, a late nineteenth-century and early twentieth-century French salonnière with a network of elites, aristocrats, writers, and intellectuals. Not all of the attendees of her salon have been documented, but this list of individuals casts a broad net over the types of people who attended her salon. We can use this small network to identify others who held their own salons as well as famous writers, painters, aristocrats, politicians, etc. We cannot, however, be confident that this group accurately represents the real demographics of Mme Greffulhe's salon, since famous or well connected individuals are more likely to be represented in the data in the first place.

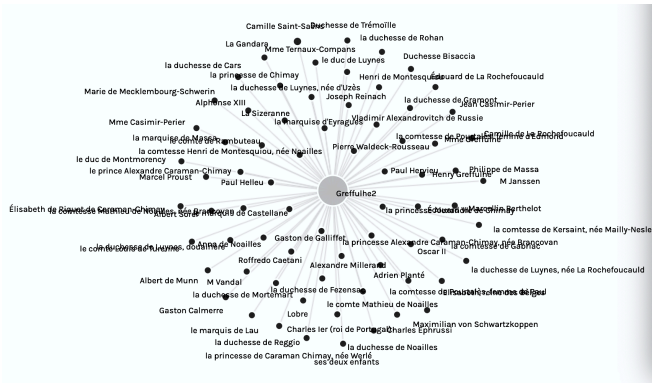

Figure 1: Ego network of French socialite Élisabeth Greffulhe within the 19th-Century Networks Database.

Figure 2: Ego network of French novelist and playwright Alexandre Dumas, père within the 19th-Century Networks Database.

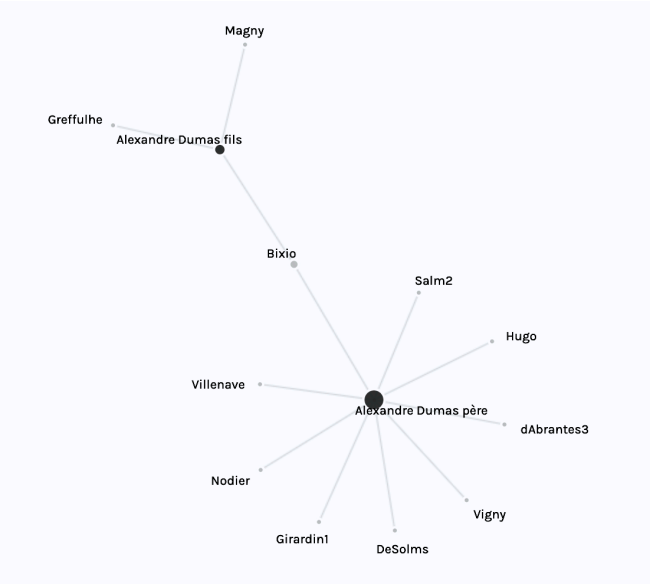

Figure 3: Linked Ego Networks of Alexandre Dumas, père et fils, within the 19th-Century Networks Database.

We can also use linked ego networks to find connections between social circles. Figure 2 shows the ego network of French novelist and playwright Alexandre Dumas, père, who attended many salons, dinners, and cénacles. Within this dataset, Alexandre Dumas, père, and Alexandre Dumas, fils, are only connected through the long-running Bixio dinner, as shown in Figure 3. Obviously, the father and son knew each other through many other venues, but the only "society" connection that I have been able to document is common attendance at the dinner of Franco-Italian politician and friend of Dumas père, Jacques Alexandre Bixio, which ran from 1856 to 1913, in some form, later led by other hosts and in other places, such as the restaurant of Paul Brébant. This social circle contained only 20 members at any one time but often serves as a connector within my dataset, due to the length of time that this dinner was held.

These limitations of the data are not apparent in the network graphs themselves. However, a subject expert would be aware of many unrepresented connections (such as the family and romantic connections missing in my data), the gaps in the data (such as missing attendees of salons), and the reasons why some networks are bridges or connections to others (such as events that are very large or held over long period of time).

4 Solutions

Aside from being aware of the characteristics of underlying data and of ego networks generally, it is also possible to create alternative visualizations. We can compare small multiples like the ones that I have shown by using grid charts. We can use tree-ring layouts (Farrugia et al., 2011).

We can also experiment with modifying the network to see what it would look like without a particular individual or connections. This solution is referred to within network visualization as "proturbing the network," suppressing nodes or edges. The idea of proturbing the network comes from robustness analysis of computer networks,(Gao

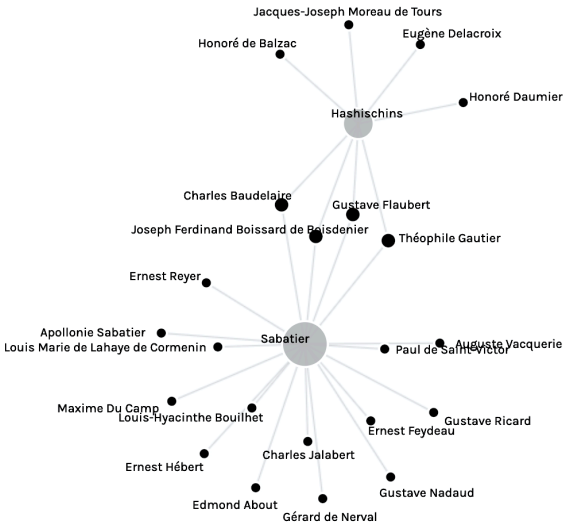

Figure 4: Linked Networks of Mme Sabatier and the Club de Hashischins within the 19th-Century Networks Database.

et al., 2012) but has also been employed in network analysis more generally (Karrer et al., 2008), but less so in humanistic network analysis. For example, removing some of the most dominant ego networks, such as Alexandre Dumas père, the comtesse de Greffulhe, or, indeed, Victor Hugo, who appears so frequently and has such high degree, that he dominates any network into which he is introduced as a node. By asking "what if" questions, such as "what would our network look like without a particular individual?" or "if two people had not met, how would this network look different?," we can perhaps see some of the connectors on whom the total network relies. Figure 4 shows the connections between the salon of Mme Sabatier and the Club de Hashischins. We can see that the connections are fairly robust since removing any of the four connectors (Flaubert, Baudelaire, Gautier, Boissard de Boisdenier) would not disconnect the two groups, whereas if either Dumas had not attended the Bixio dinner, they would not have been linked in the dataset.

Finally, we can also create "super-nodes," or new communities, to reveal other important connectors within the data. I will demonstrate this last experiment as my current preferred solution for seeing the ways that ego networks - particularly of high degree individuals - distort networks comprised of mostly smaller ego networks. By creating communities of lower degree individuals (in my data often journalists, translators, or family members of writers), the relations between the "big fish" and the schools of fish that swim around them can be more fully articulated. That said, the goal of removing and adding the most high degree nodes, or of creating new communities out of lower degree nodes, is not to access the "true" structure of the network but to see the relations between different parts of the network, independent of the weight of individuals or individual connections.

References

Comsa, M. T., M. Conroy, D. Edelstein, C. S. Edmondson, and C. Willan (2016). The french enlightenment network. The Journal of Modern History 88(3), 495–534.

Conroy, M. and K. Elo (2020). Picturing the politics of resistance. Digital Histories, 221.

Edelstein, D., P. Findlen, G. Ceserani, C. Winterer, and N. Coleman (2017). Historical research in a digital age: Reflections from the mapping the republic of letters project. The American Historical Review 122(2), 400–424.

Farrugia, M., N. Hurley, and A. Quigley (2011). Exploring temporal ego networks using small multiples and tree-ring layouts. Proc. ACHI 2011, 23–28.

Gao, Z., Z. Gu, and W. Wang (2012). Spsi: A hybrid super-node election method based on information theory. In 2012 14th International Conference on Advanced Communication Technology (ICACT), pp. 1076–1081. IEEE.

Karrer, B., E. Levina, and M. E. Newman (2008). Robustness of community structure in networks. Physical review E 77(4), 046119.

Wang, X., Y. Ran, and T. Jia (2020). Measuring similarity in co-occurrence data using ego-networks. Chaos: An Interdisciplinary Journal of Nonlinear Science 30(1), 013101.