# “Silicon Valley Mafia”: Network Analysis on Startups and Venture Capitalists

Ali Alkhatib, Chentai Kao
{al2,chentai}@stanford.edu

Final Report

## Introduction

Startup culture and entrepreneurship have earned enormous interest among researchers from industry and various fields of academia, but relatively little attention has been paid to the investors from whom the capital for these startups and entrepreneurial endeavors originate. These investors, largely Venture Capitalists (VCs), represent a unique but not-well-known community of people; their opacity has eluded critical exploration from researchers across disciplines. The clandestine nature of these people was partly due to the challenges researchers faced in their efforts to access data relevant to these people. Recently, the emergence of accessible, consumable data illustrating the relationships between investors and entrepreneurs - or more accurately startups - has made the study of VCs more practical than it once was. Using a number of tools native to network analysis, we yield the first insights from an exploratory study of the investors of Silicon Valley startups, especially with an eye towards potential indicators of communities, cliques, and other in-group organizational structures.

## Prior work

Some particularly relevant research has explored topics similar to or overlapping with this research. Bygrave [[1],[2]] studies the co-investment network in Venture Capitalists’ (VC) portfolio companies, the spreading of financial risk, and the sharing of knowledge. These studies identify the existence of cliques among top VC firms, mostly based in California, though it found no evidence that these VCs excluded other VCs from their co-investments.

This research argues that the degree of co-investment reflects the degree of uncertainty, estimably related to financial risk. With this being said, this researcher doesn’t adequately explore the detail of economic behavior of co-investment for our purposes; we can investigate this by examining detailed transaction data associated with each investment thanks to the cumulative efforts of members of the community of entrepreneurs, journalists, and the effects of legal requirements dictating some level of disclosure. Finally, Bygrave proposes a general model for co-investment network, which gives us inspiration regarding what factors should be considered in order to model co-investment networks.

While Bygrave explores the relationships of investors among one another, Antonio et al. examine the association between VC presence and employee growth in startups, thus focusing on “investor-startup” relationships [[3]]. This study investigates the impact of VC financing events upon the growth of startups. The dataset used in this study documented the relationship between the growth of employee over successive rounds of financing. This study leveraged data spanning 1994 to 2000, including hundreds

[1] Bygrave, William D. "The structure of the investment networks of venture capital firms." *Journal of Business Venturing* 3.2 (1988): 137-157.

[2] Bygrave, William D. "Syndicated investments by venture capital firms: A networking perspective." *Journal of Business Venturing* 2.2 (1987): 139-154.

[3] Davila, Antonio, George Foster, and Mahendra Gupta. "Venture capital financing and the growth of startup firms." *Journal of business venturing* 18.6 (2003): 689-708.

of Silicon Valley-based companies.

Bygrave and Antonio et al. illuminate the research area which we seek to illuminate further, but we turn to other researchers for inspiration regarding methodology. Cha et al. study Twitter users in the context of the directed network of followers (those following a user) and “friends” (those a user is following), attempting to determine underlying characteristics of influential users [[4]]. The intuitive perspective Cha et al. critique is the notion that users with the most followers are necessarily the most influential, which they point out is a difficult question to pose meaningfully due to the inherent challenges of quantifying an individual’s influence and the fact that experimental verification of findings made from observational research is impossible in that lab settings cannot replicate or emulate the nuanced relationships found in real-world settings.

The researchers consider an alternative theory, essentially that there are no hyper-influential users driving trends, and identifies numerous additional questions and ambiguities surrounding what we know about influence in communities: How does an individual’s expertise in a given area affect their influence? Does that aspect of an individual - their level, let alone area, of expertise - itself influence their ability to effect change in others’ behavior? The researchers turn to Twitter to explore various metrics for influence, formulating an empirical study rather than an experimental one.

Cha et al. make interesting discoveries which challenge simplistic intuition about influence in a network. First and foremost, of the three criteria conventionally used to represent influence (in-degree, retweets, and mentions), the groups representing high-scoring users “have little overlap”. Instead of substantially similar groups, high-scoring individuals in each group attain their respective statuses for different reasons - highly retweeted users have high “content value”, whereas users with high in-degree scores are simply very popular. These findings quantitatively confirm and substantiate until-then anecdotal arguments that in-degree and influence are not necessarily strongly and positively correlated.

Specifically, Cha et al. critique “the traditional influentials theory” on the following bases:

1. That network simulations faithfully replicating the free-flowing exchange of information demonstrate that influential users do not initiate enough exchanges to explain all diffusions of information, suggesting that the theory itself is flawed; and
2. that the Internet - with its affordances for decentralized communication and information-exchange - problematize the theory that highly central influencers act as gatekeepers of substantial information diffusion.

This research finds that measures such as in-degree indeed mislead one’s estimate of a user’s influence, leading to differences in overlap between the users with high in-degree values and the users with high retweet proportions and mentions. These critical approaches to measuring an individual’s influence suggest ways that we might approach the analysis of people’s influence in other fields and settings than Twitter.

Cha et al. raise important points regarding the study and measure of influence through Twitter, but several considerations are worth pointing out. First, none of the metrics which they measure as abstractions of influence take into account the limited overall influence that people can exert upon one another. Put more directly, not everyone can have maximal influence over everyone else. In each of the

[4] Cha, Meeyoung et al. "Measuring User Influence in Twitter: The Million Follower Fallacy." *ICWSM* 10 (2010): 10-17.

measures the researchers use, there is no structure enforcing or emulating this dynamic.

To use an example, in theory it is possible for every user on Twitter to follow every other user on the network. In reality, some nominal limit is imposed, but it is by no means an imposition on normal users which dictates a scarcity mindset. As a result of this disconnect, people may have measures suggesting influence which inflate their actual influence because in the “real world” people’s capacity to be influenced by others is limited by constraints such as finances.

This weakness is also applicable, in more limited forms, to their measures of influence employing retweet metrics and mention metrics (excessive retweeting may cause one’s own influence to suffer, putting resistant pressure on retweeting; similarly, excessive mentions may frustrate mutual followers). This case is empirically found in many “joke” accounts on Twitter, which the researchers do not address.

Despite these limitations, we feel that the general approach of using a character’s finite resources - in our case, finances - to determine the weight of a relationship can bear significant insight into the relationships of investors among one another as well as with companies. As we discuss later, our initial findings seem to corroborate this hunch.

One question remains, and that is the criteria by which we evaluate the algorithm which analyzes investor performance. While superficially this would seem like a shallow problem, the disparate goals of VC firms [[5]] problematizes a monotonic evaluation of VC performance according to a criterion such as acquisition rate or current valuation, to say nothing of the challenges associated with finding current valuations of non-public companies, a category in which most Silicon Valley companies fall. Nevertheless, we can safely conclude that VC firms generally aspire toward the same goal - the accumulation of wealth - and various indicators such as acquisition and operating status indicate, or at least imply, the VC firm’s performance.

## Data collection process

We use data from Crunchbase, an online database of investment record on startups and investors. Instead of calling upon their API innumerable times, we leveraged a “data-dump” published by Crunchbase which contains all records up to the current month and year. We focus on investment records, which include investor, company of interest, time of investment, and magnitude of investment. By exporting the data to a more universally accessible CSV format, we are able to parse it and build a graph with the Python SNAP.py library.

Since we also looked at geographic investment relationships, precise locations of companies and investors proved necessary. Fortunately, even this data was, at least partially, available in the data dump provided by Crunchbase. We utilized Google’s Geolocation API, which returns the query’s location represented by latitude and longitude, allowing us to turn qualitative values into quantified, measurable, comparable distances and nodes. We store location for all companies and investors in a single file to avoid calling the online API for each data access. This avoids exceeding the hourly, daily, and monthly API call limits.

Currently we use three kinds of edge weights: investment amount, distance between investor and

[5] "Assessing Fund Performance: - Silicon Valley Bank." 2014. 13 Nov. 2014 <https://www.svb.com/publications/industry-trends/venture-capital-update/assessing-fund-performance--using-benchmarks-in-venture-capital-(pdf)/>

company, and time of investment. We hypothesized that the timestamp might provide important information on the investment itself, where recency positively relates with relationship strength. Additionally, we wanted to detect the change of investment behavior over time. For example, if investments plateaued or declined over time, that might indicate something different from increasing investments.

## Mathematical background

Our data proved to be fairly rich. With nearly 45,000 nodes and more than 79,000 edges, we could apply a number of algorithmic approaches to analyzing this data, once cleaned and normalized properly.

To find geographic relationships, we plot a histogram of distances between companies and VCs behind each corresponding company's investment. This would shed light on top VCs preference on nearby companies, as well as different investing behavior between VC and personal investor. In order to determine the distances between two locations, we use the Haversine formula to determine the distance between two points on a spherical object, in this case the Earth. Specifically, given that

$$cos(c) \ = \ cos(a) * cos(b) + sin(a) * sin(b) * sin(C)$$

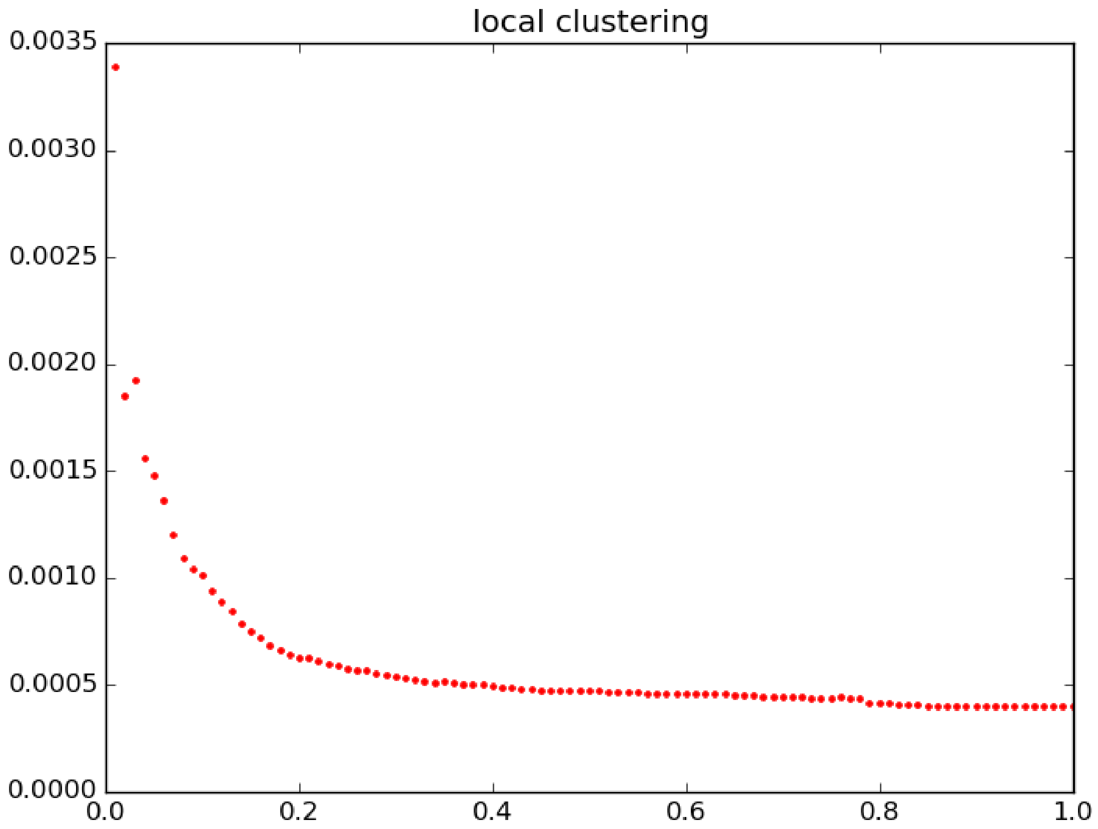
where *a*, *b*, and *c* represent the distances between 3 points on a spherical graph - in this case, the Earth - and *C* represents the angle opposite line *c*, we can calculate the distances between investors and startups, given their geographic coordinates.

In addition, we investigate whether investment nodes follow a power law. Our approach in this case involves plotting a histogram of node degree and fitting a power function to the results. We can also examine power law for other quantities, such as investment amount, one characteristic which can inform or determine edge weight.

## Results

We considered the network according to several models to allow us to evaluate the nodes and their relationships in various ways, potentially yielding new insights with each approach. Ultimately, we found that this network most conformed to a bipartite graph model, however we found other approaches informative and indeed found a typical bipartite model an imperfect description of the network itself.

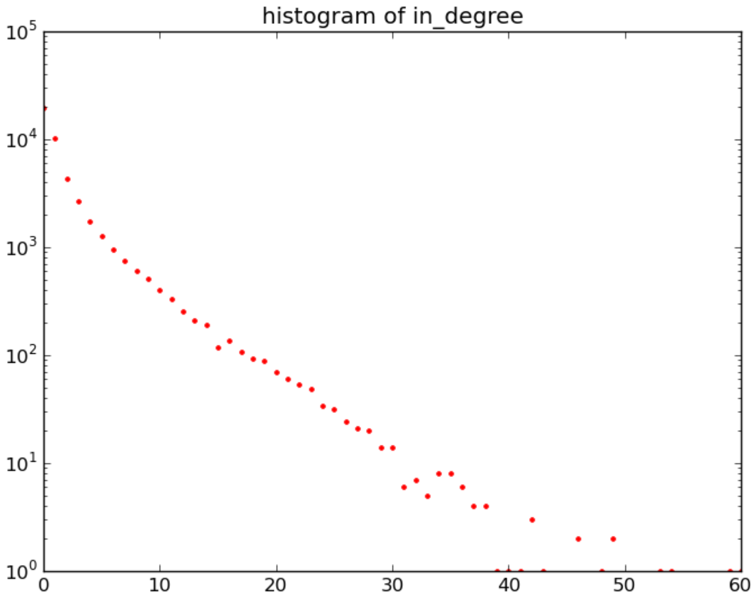
With more than 44,831 nodes and 97,301 edges, this network is very sparsely connected, though as we will find there is a significantly more densely connected core with disproportionately more edges than the average path of 1 suggests. Indeed, with a graph diameter of 4 when evaluated as an undirected, unweighted graph, this cluster of nodes seems to confirm the intuition with which we entered - that investment in Silicon Valley is dominated by a relatively small cohort of influential and somewhat like-minded VC firms, as evidenced by their interconnectedness, illustrating their investment and co-investment patterns.



*Local clustering coefficient plot; y-axis represents the clustering coefficient, x-axis 0.01 steps of number of vertices*

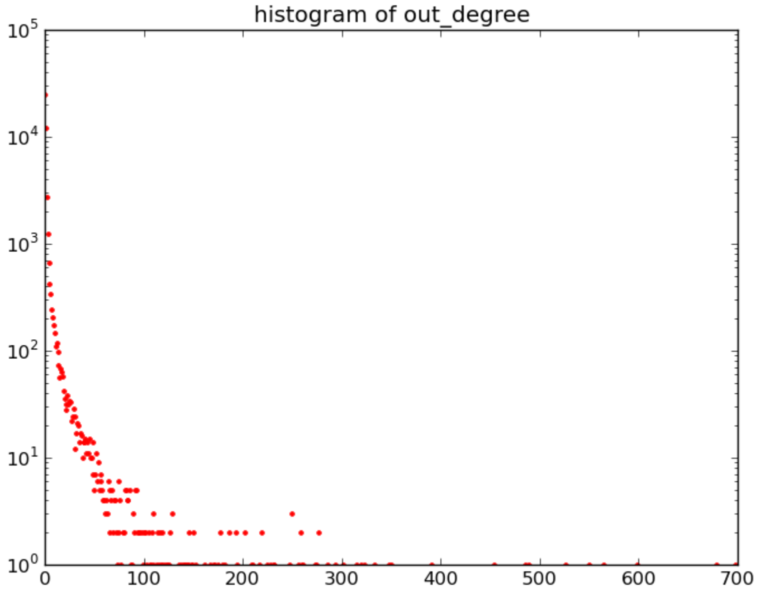
The clustering coefficient illustrates a similar conclusion in a different way; as we see from this graph, the clustering coefficient drops logarithmically as more of the network's nodes are included. At the far left, with the top 0.001 (or top 0.1%) nodes as measured by degrees, we see a clustering coefficient of nearly 0.0035. This clustering coefficient drops sharply and settles near a clustering coefficient of 0.0005 by the time one includes the top 20% of nodes. The average clustering coefficient is 0.0004009.

Considering the investment network as a bipartite graph provided some interesting techniques for analysis. Specifically, evaluating the in-degree distribution of startups as well as the out-degree distribution of VCs provided different insights despite similar methods. In particular, we found that the in-degree distribution of startups in our network, illustrating the number of VCs startups tended to successfully persuade to invest in them, declines logarithmically; the vast majority of startups only win the investment of a small handful of companies, with 10^4 startups soliciting the investment of only one VC. Similarly few startups raise funds from as many as 40 or more VC firms.



*Histogram of in-degree of startups; x-axis represents count of investors, y-axis count in logarithmic scale*

The out-degree distribution of investors tells the same story from the other side, interestingly revealing different details and a power law distribution we did not anticipate finding. Specifically, the number of VCs which invest in many companies, represented at the bottom right of the graph, is vanishingly small. Instead, most VCs tend to invest in fewer than 100 companies, represented by the sharp slope of the histogram as x (number of companies invested in, or *out-degree*) approaches 1.

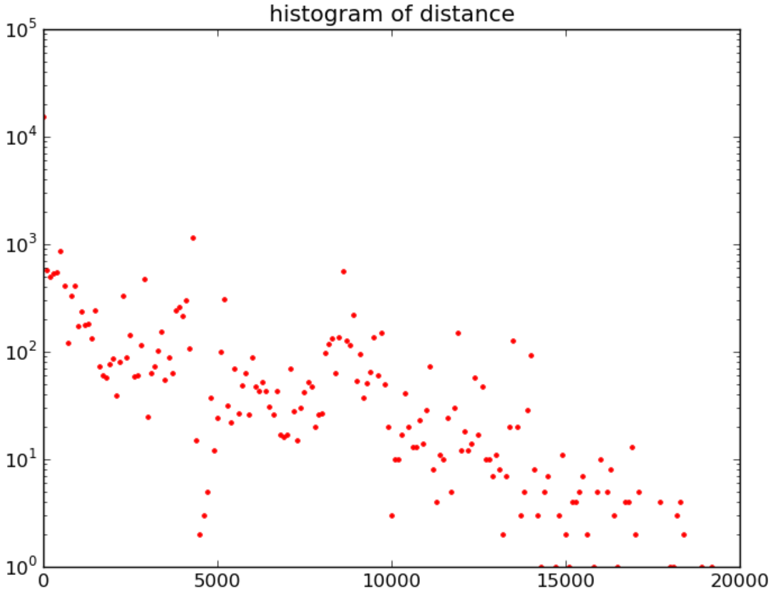


*Histogram of out-degree of VCs; x-axis represents count of investments, y-axis count in logarithmic scale*

Together, these figures tell two complementary stories. From the startup's perspective, having more than 10, 20, or 30 VC investors is an increasingly rare proposition to the point that only a small handful of startups even claim more than 30 co-investors. Investors, for their parts, tend to invest in fewer than 100 startups and predominantly invest in fewer than 50 startups, but a number of VC firms invest in as many as 200 or more startups.

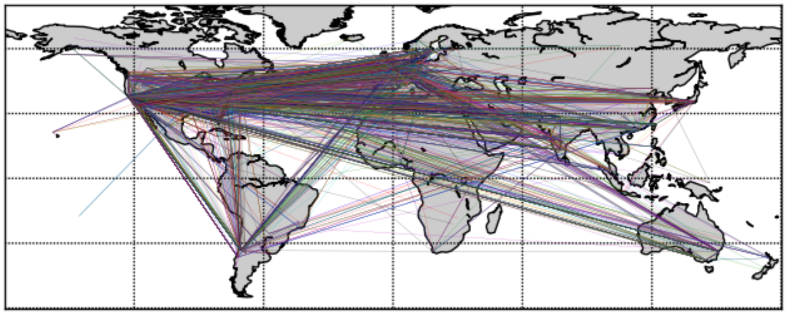
Finding that there are more VC firms investing in many startups than there are startups soliciting investment from many investors corroborates the intuition of the difficulty of raising capital for a startup, even with VCs who invest in many different startups or indeed in "serial entrepreneurs", who may create several startups over their career and seek investment on multiple occasions.

We considered that the distances between investors and the startups in which they invest might represent a confounding variable. The analysis of distances between investors and the startups in which they invest suggested a stronger preference for more local startups, but as before we found that VCs were not as averse to investing in distant startups as we assumed they would be; a significant minority of VCs invested in startups more than 10,000 kilometers from their primary locations.

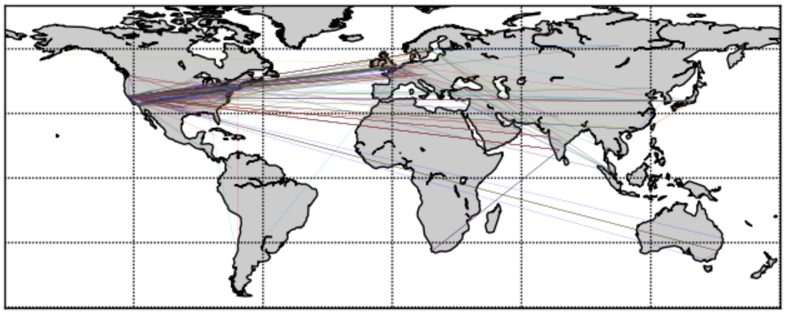


*Histogram of distances from VCs to their startups; x-axis is distance in kilometers, y-axis is count in logarithmic scale*

We also considered that there might be a noticeable trend toward a single geographic region where money tends to funnel into or out of certain places. Again intuitively we expected to see edges, when nodes were plotted on a geographic projection (e.g. Mercator), would visibly tend to point toward Silicon Valley, although we had no intuitive guess regarding where the nodes would originate, except that they might generally be local. Plotting along these lines revealed not only a hub in Silicon Valley, but also hubs in Europe, India, Japan, and South Africa.



*Undirected network graph of all investments, where nodes are plotted at their associated geographic coordinates*



*Undirected network graph of top 90% of investments by scale of investment*

## Discussion

Our findings confirm some of the intuition we came into our research with, while confounding others. While we indeed found a very densely connected network of the top 50-100 venture capitalist firms quantitatively identifiable from their investment behavior across various criteria such as investment magnitude and frequency. These findings tend to corroborate rankings made intuitively and without the backing of data of this nature. The asymmetric structuring of relationships between VCs and startups, characterized by VCs' increased likelihood to invest in as many as hundreds of startups, while startups are significantly less likely to acquire funding from as many investors.

Findings such as these suggest that our quantitative approaches to evaluating VCs bear some resemblance to metrics such as return on investment. Crucially, our approach differs from other metrics such as ROI in that ours evaluates behavior rather than result. While we cannot conclude that this approach can reliably predict VC performance *a priori*, we feel that this represents a promising first step in modelling expected VC performance.

More to the original goals of this research, we indeed found that the core of investors who co-invest with one another form a very tightly connected graph network with a small diameter; one might infer from this that popular startups tend to attract many investors, but the previously discussed statistics of the network - especially that startups tend not to attract more than 30 investors - suggest instead that these investors are collaborating on many startup endeavors rather than on a single or small handful of startups.

## Future work

Our analysis only superficially outlined the different nature of investor behavior among top VCs compared to others, and further analysis of the collaboration network represented by co-investment of these VCs could identify characteristics we decided not to seek from the outset. For instance, whether certain cohorts of VCs invest together but not with others within this more densely connected network (ie some sort of faction or tribe behavior), or identifying emergent preferences for startups (according to traits, relationships, or something else) by potential investors. Further exploring the behavior and identifying patterns in investments among VCs in this more densely connected core might yield further insights. For example, we have no insight as to whether the geographic distribution of investments changes temporally or not; a shift in investment over time might reveal more than we uncovered, but this avenue was outside of the scope of our research.

In our analysis we encountered entities which behaved both as startups and as investors. To be more specific, we found that some startups (e.g. Reddit) had itself made investments into other startups. In our analyses we considered these cases as two nodes (one as an investor and an identically named node as a startup), which enabled us to continue our analysis without further complications. These occurrences, though rare (representing fewer than 1% of all nodes), nevertheless represent a rich area to explore further. Aside from the uniqueness of these nodes, some of these entities were found among the "top 100" of either VC firms or startups.

Finally, while we examined the relationships of investors to the startups in which they invested, we notably did not have access to data suggesting what we would consider "negative" relationships. In other words, opportunities to invest in startups that VCs actively rejected. The phenomenon of a VC passing on an investment is indeed more common than successfully raising money among startups, suggesting that there are dramatically more edges that could populate a network graph of investors and startups. While this would be an interesting subject to pursue, no comprehensive record of such events occurs; VCs may keep records of investments they decline to make, but no VCs offer this data to the public (let alone submit this data to a comprehensive database for inclusion in analysis such as this). This data would potentially add a dimension to the network that would expose relationships in greater quantity and quality.