The Social Network of the Planetary Data System: A Comparative Analysis of Network Representations

Mark Avnet
Kate Martin

Supervisor:
Daniel Whitney

May 16, 2006
ESD.342 Advanced System Architecture

1. Introduction

Since the beginning of the space program, the National Aeronautics and Space Administration (NASA) has collected several terabytes of data about our solar system. In fact, the amount of data available far exceeds the combined analytical capacity of the entire planetary science community. As part of an effort to mitigate this problem, NASA has created a data management tool called the Planetary Data System (PDS).[1] The PDS, established in the early 1990s, is an arrangement of eight nodes located at various NASA Centers, universities, and research institutions across the country. Although the use of the PDS is currently somewhat limited, the intent is that it eventually will be the primary repository of all planetary data. Therefore, analyzing scientists' usage of the PDS can provide some important insights about the planetary science community. For example, answering a few key questions could help the PDS system operators understand why the system is not used to its full potential. One such question is: does the categorization of the data match the research communities that actually form around those data?

One purpose of this project is to develop a better understanding of the usage of the Planetary Data System. The system is used in two distinct but complementary ways. The first is that the collectors of data can upload their data to the system to, in a sense, "immortalize" it (a few people associated with PDS indicated in informal interviews that it was important to have a system like PDS because the datasets can last forever, but people do not). The second type of system usage occurs when scientists or others

[1] All background information about the Planetary Data System in this section from: National Aeronautics and Space Administration, "About PDS," *Planetary Data System*, <http://pds.jpl.nasa.gov/aboutpds.html>, accessed on May 11, 2006.

download the data for research or other purposes. This study focuses on the first type of usage because data for the second type were sparse.

To study patterns of collaboration on the usage of the PDS, several representations of the affiliation network are used. In the representation most typical to the social network literature, a network was created with authors (of datasets) as nodes and datasets as the edges. Other information about the PDS and the datasets are used to compare several representations of essentially the same network of author collaborations and planetary data. This information is also used to study the community structure and to compare community algorithm results and centrality measures to contextual understanding of the author roles and interactions. The use of multiple network representations allows us to fulfill a second purpose – to examine some of the limitations of popular network analysis techniques and metrics. We examine the basic statistics of each representation of the network and consider the effects of representation and weighting on the results. The paper concludes with some suggestions for future work in both of these areas.

1.1. PDS Structure

The conceptual structure of the PDS is involves a central Project Management office and eight separate nodes on which the data are contained. The term "PDS node" used here is quite distinct from the network nodes discussed in this paper. Furthermore, the PDS nodes are not just the servers on which the data are contained. The term actually refers to "teams" of people that span research centers and universities. The five science nodes, named for five major sub-disciplines of planetary science, are Atmospheres, Geosciences, Planetary Plasma Interactions (PPI), Rings, and Small Bodies. The three remaining nodes provide support. These nodes are Engineering, which "provides systems engineering support to the entire project;" Navigational Ancillary Information Facility (NAIF), which "supplies calibration and ephemeris information;" and Imaging, which "offers expertise in sophisticated image processing." Figure 1 shows the conceptual layout of the PDS.

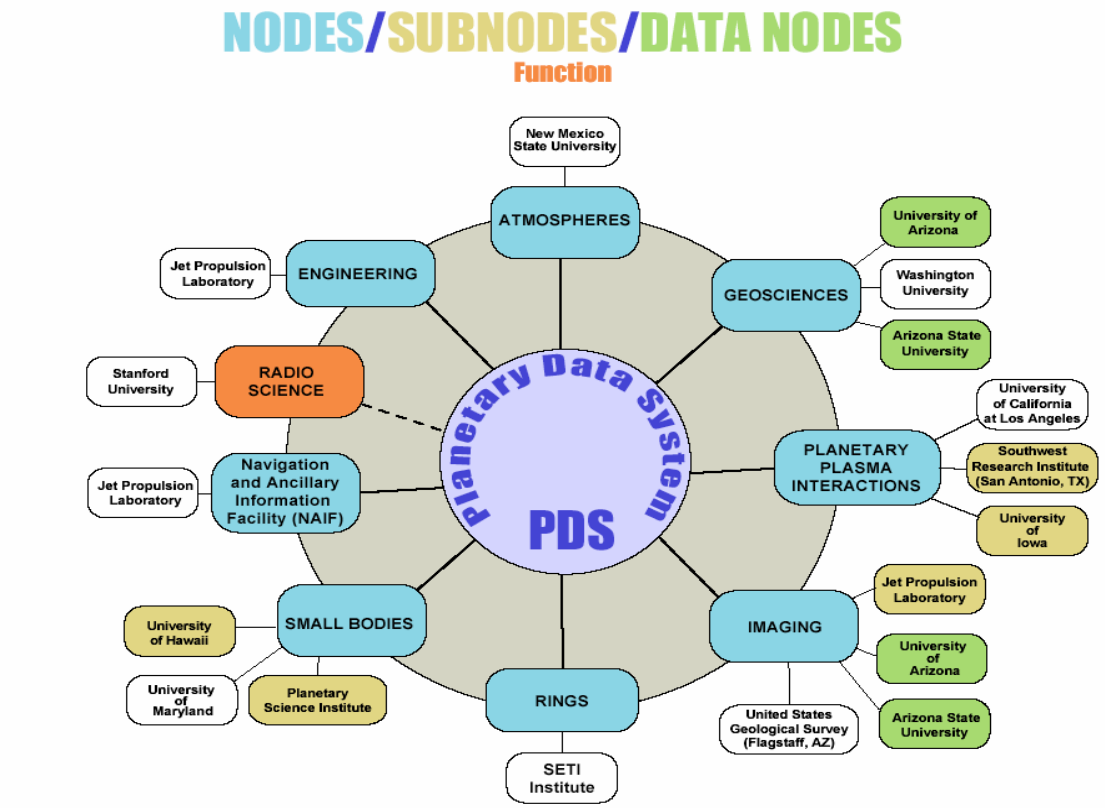

Figure 1. The conceptual structure of the Planetary Data System.
Image courtesy of NASA. *Source:* http://pds.jpl.nasa.gov/files/node.pdf.

The Project Management function of the PDS is located at NASA Goddard Space Flight Center (GSFC) in the Solar System Exploration Data Services Office. This, however, implies a level of centralized control that does not exist in the actual implementation of the system. Each of the nodes run independently of the others, and the use of each node is dependent on the people managing it. The leadership of the nodes changes every five years (the third selection was completed in June 2004) via a competitive grant solicitation process run by NASA. These changes in PDS node leadership can affect activity on the node. In some cases, the universities desire management responsibility for a node not because of the scientific importance of the PDS but because of political pressure and promise of funding for related programs. In many ways, the PDS is as much a public relations tool as a data management tool. Therefore, the actual structure of the PDS that has arisen from the loose-knit usage of the system is rather different from the prescribed structure shown in Figure 1.

2. Network Representation

Much of the literature on the science of networks implies the importance of the choice of representation on the structure of the network. Collaboration and affiliation networks are often considered to fall into the broad category of "social" networks, but this is not an entirely accurate description. The relationships between the actors are more professional than social, and a relationship, once established, cannot be broken (the authors cannot "unwrite" a paper). For these reasons, Watts and Strogatz (1998) refer to the network of film actors as a "surrogate" for a social network but point out that the aforementioned characteristics provide "the advantage of being much more easily specified." Additionally, Amaral et al (2000) refer to that network as an economic one because it, in some way, represents the flow of money, at least within the film industry.

Furthermore, the meaning of the type of network can change considerably if the human actors are represented as edges and the events as nodes. In any type of collaboration network, a social network (broadly defined) surely exists among the authors that develop and submit the datasets. Still, if the nodes and edges are switched, this same network can be viewed as an information network in which the human actors serve as the channels of information flow. As we will see, in the case of the Planetary Data System, human actors can similarly represent links between technological artifacts. In this type of representation, the same network might even be seen as a technological network. The point of this line of reasoning is not to claim that humans serve as the wiring of an engineered system but merely that the way that one chooses to represent a network can have important effects on the results of the analysis. Still, surprisingly few papers actually undergo a comparative analysis of the same network represented in a variety of ways. In this paper, we compare 12 different representations of the PDS collaboration network.

2.1. Categorization of PDS Datasets

The datasets in the Planetary Data System are categorized on several levels. Each dataset is located on one of the nine nodes (including a catch-all referred to as "N/A"), and each is derived from data collected by a particular spacecraft or telescope, referred to

as an "instrument host" in the vocabulary of the PDS. A result of this structure is that the links between authors are created not just by collaborating on the creation of a dataset but also by working, perhaps independently of collaboration on the dataset, on the same PDS node or instrument host. Furthermore, PDS nodes, instrument hosts, or datasets are "connected" by the presence of common authors working on them.

2.2. Bipartite Network Representations and 1-Mode Projections

In the language of networks, this arrangement essentially means that three bipartite networks can be extracted from the database. In each of these 2-mode networks, authors represent one of type of node. The other type of node corresponds to PDS nodes, instrument hosts, and datasets for each of the three networks. It would also be possible to create other bipartite networks that do not involve authors, such as connecting datasets to instruments hosts. This type of network, however, is unlikely to be as interesting. It would not be surprising, for example, to demonstrate that a data set on some geological property of Mars is located on the Geosciences PDS Node.

With the three separate bipartite affiliation networks established, we were able to create 12 distinct 1-mode networks for comparative analysis. First, we split each of the three bipartites, creating six separate 1-mode networks: (1) authors as nodes, PDS nodes as edges; (2) PDS nodes as nodes, authors as edges; (3) authors as nodes, instrument hosts as edges; (4) instrument hosts as nodes, authors as edges; (5) authors as nodes, datasets as edges; and (6) datasets as nodes, authors as edges. For the sake of our analysis, we refer to the representation of PDS nodes and instrument hosts as network nodes as technological networks and the representation of datasets as network nodes as an information network.

In addition, each of these networks contains an implicit weighting that results from multiple collaborations between, say, two authors on several datasets or spacecraft. In reality, this weighting is an intrinsic property of nearly all collaboration and affiliation networks. Newman et al (2001) have done some analysis of bipartite networks without making the simplification to a single mode, and Marchiori and Latora (2000) have even made an attempt to quantify the strength of connections in a metric that they call

connectivity length. Still, the implicit weighting resulting from multiple nodes of one mode between two nodes of the other mode is ignored in most of the literature.

In our analysis, we decided to compare the weighted and unweighted versions of the six PDS networks. To do this, we created six additional networks by replacing each of the non-zero values with 1s in each of the weighted networks to create an unweighted network corresponding to each. In addition, like Marchiori and Latora, we consider the entirety of the networks in most of our analysis (except where this is not possible) instead of just the giant connected component. The process of splitting the PDS database into 12 weighted and unweighted 1-mode networks is represented in Figure 2.

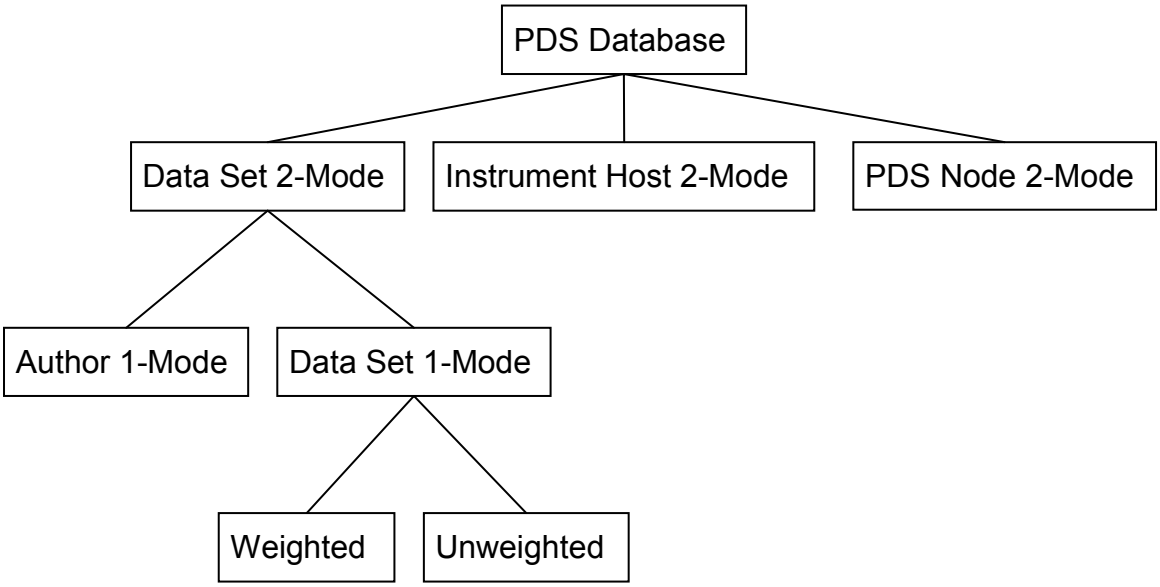

Figure 2. Hierarchical structure of the PDS network representations. The complete hierarchy is shown for only one of the 1-mode representations and its weighted and unweighted versions to maintain clarity of presentation, but the presence of the other 10 bottom-row elements (for a total of 12 networks used in the study) is implied.

3. Network Statistics

In this section, we examine some of the basic statistics about the PDS network. These statistics are presented in a table of the form presented in Mark Newman's review paper (2003). A discussion of the columns of Table I are in Sections 3.1 through 3.4.[2]

Table 1. Table of the 12 networks extracted from the three bipartite networks of the PDS database. This table is of the form of Table II in Mark Newman's review paper (2003). Instead of categorizing the networks as social, information, technological, or biological, we separate them by representation – whether authors or events are the nodes. We omitted the type of the network since all are undirected, and we also omit one of the measures of clustering coefficient because our tools only computed the other. We have added to this table the equivalent random network approximations for path length and clustering coefficient in the column following each of those metrics.

Type of Node	Network	Weighted?	n	m	$\langle k \rangle$	l	$\log n / \log \langle k \rangle$	α	$C^{(2)}$	$\langle k \rangle / n$	r
Authors as Nodes	PDS Nodes	No	439	27493	125.3	1.886	1.260	-0.31	0.981	0.285	0.80
	PDS Nodes	Yes	439	27527	125.4	1.886	1.259	-0.31	0.988	0.286	0.80
	Instrument Hosts	No	439	8240	37.5	2.703	1.678	-0.49	0.929	0.086	0.68
	Instrument Hosts	Yes	439	8581	39.1	2.703	1.660	-0.48	1.092	0.089	0.55
	Data Sets	No	439	3240	14.8	3.1	2.260	-0.60	0.936	0.034	0.96
	Data Sets	Yes	439	4366	19.9	3.1	2.035	-0.53	1.534	0.045	0.95
Events as Nodes	PDS Nodes	No	9	11	2.4	1.861	2.458	-1.1	0.62	0.272	-0.70
	PDS Nodes	Yes	9	28	6.2	1.861	1.202	-0.73	1.927	0.691	-0.52
	Instrument Hosts	No	103	282	5.5	2.426	2.726	--	0.715	0.053	0.039
	Instrument Hosts	Yes	103	520	10.1	2.426	2.004	--	1.675	0.098	-0.042
	Data Sets	No	1046	5820	11.1	2.761	2.886	-0.27	0.937	0.011	0.99
	Data Sets	Yes	1046	6514	12.5	2.761	2.757	-0.29	1.455	0.012	0.97

3.1. Network Size

The networks resulting from our analysis of the PDS are relatively small in size. The PDS contains 1,046 datasets with data from 103 instrument hosts on nine nodes of the system (including the N/A node designation). The datasets have a total of 439 authors. Of course, the number of nodes is not affected by weighting, but the number of edges is affected because weighting is represented as a number greater than 1 in the adjacency matrix. In these networks, an edge has a value of 1 if it represents, for example, one common data set between authors or one common author between instrument hosts. An edge representing two such links in common would have a value of 2 in the adjacency matrix, and so on. Therefore, a weighted edge also could be drawn as multiple edges between two nodes in the graph. In this way, *m* does represent the actual

[2] Many of these metrics were calculated with UCINET (Borgatti, 2002).

number of edges in the network. The effect on <*k*> of the weighting simply is a result of the effect on *m*.

3.2. Path Length and Clustering

Each of the 12 networks has a relatively short path length and extremely high clustering coefficient. High clustering is not at all surprising for 1-mode projections of bipartite networks. In a bipartite network, each type of node can connect only to a node of the other type. Thus, each edge between nodes of one type must be channeled through a node of the other type. Thus, if multiple nodes of the former type are channeled through the same node of the latter type, clusters (as represented by triangle motifs) will necessarily arise in the 1-mode projection. This is shown in Figure 3 conceptually and in Figure 4 for the PDS network of authors and datasets. This high clustering, in turn, contributes to shorter path lengths.

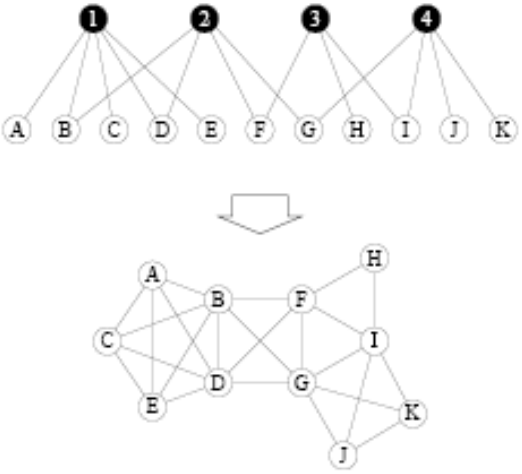

Figure 3. This simple projection of a bipartite network onto a 1-mode network demonstrates the high clustering in the projection. Also note that another 1-mode consisting of nodes 1, 2, 3, and 4 also could be created. In that network, a triangle would exist between 2, 3, and 4, but 1 would be connected only to 2. Thus, that 1-mode would have a lower clustering coefficient than the one shown here. *Source:* Newman et al (2001)

Still, it is not entirely meaningful to discuss path length and clustering without considering those metrics as compared to the expected values for a random network. For this reason, these metrics are presented along with the actual values. Watts and Strogatz (1998) use these random network approximations to determine whether a network can be said to be a small world network. They define the small world phenomenon as $l \underset{\sim}{>} l_{random}$

but $C >> C_{random}$. The second condition is clearly satisfied by all 12 of the networks with the possible exception of the unweighted network of PDS nodes connected by authors as edges, for which $C = 0.62$ and $C_{random} = 0.272$.

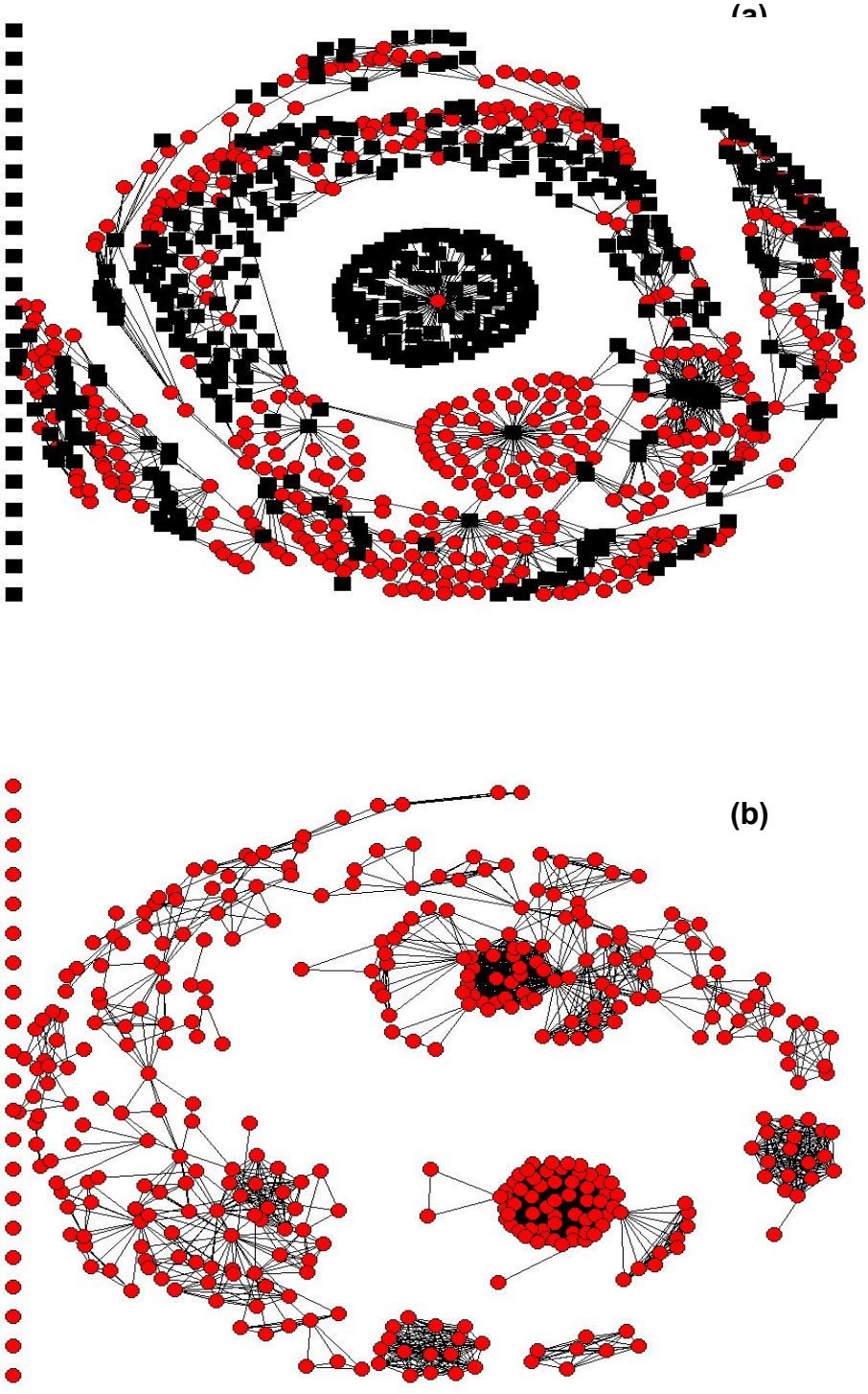

Figure 4. Graphical representation of the bipartite network of authors and datasets. (a) The bipartite network connecting authors (red circles) to datasets (black squares). (b) The corresponding 1-mode social network with authors as nodes and datasets as edges. Note the high frequency of triangle motifs and, therefore, high clustering in the unipartite network.

The first condition, however, is a bit more ambiguous. For both the weighted and unweighted versions of each of the "social" networks, $l > l_{random}$, but l is a good deal greater than l_{random} in some of the cases. In both "technological" networks (PDS nodes and instrument hosts as nodes), weighting makes a significant difference for this small-world condition. To understand this, first note that path length is independent of any weighting on the edges.[3] This can be illustrated by thinking of a weighted edge as multiple edges between two nodes: when determining path length, one need only use one of the edges. The definition of $l_{random,}$ however, does depend on the weighting of the edges. Because of this effect, $l > l_{random}$ for the weighted networks, but $l < l_{random}$ for the unweighted ones. As for the "information" network, $l < l_{random}$ in the unweighted case, though only by a small amount. In the weighted "information" network, however, this small-world condition appears to be met: $l \gtrsim l_{random}$.

Although the inclusion of weighted values for clustering and path length can give a more complete picture of the network, it also can obscure the results and perhaps even render these metrics meaningless. Indeed, one must be somewhat skeptical of the clustering coefficients for the weighted networks since most of them are greater than 1. Mathematically this result makes sense. If a weighted edge is represented as an additional edge between the same pair of nodes, weighted edges in triangles could substantially increase the number of triangles relative to the number of connected triples. Still, this effect means that the clustering coefficient is not normalized. Similarly, the result that weighting does not affect path length even while it does change the expected path length of the equivalent random network also should be cause for suspicion. Undoubtedly, more work needs to be done on the analysis of weighted networks. Marchiori and Latora's connectivity length (2000) provides a possible solution to some of these problems, but it is not at all clear that this metric can resolve the entire issue.

[3] According to some measures, including connectivity length (Marchiori and Latora 2000), weighting can be thought of as a length. In the representations used in this study, however, the lengths of all edges are equal, and weighting is taken to be equivalent to having additional edges.

3.3. Degree Distributions

3.3.1. Features of the Degree Distributions

Degree distributions are a popular means by which to understand the likelihood that a node selected at random will have a high or low degree. Degree distributions are one way that real graphs differ dramatically from random graphs, at least of the simple type like those studied by Erdös and Rényi (1959). When edges in a graph are present or absent with equal probability, the degree distribution of the graph is a Poisson or binomial depending on its size. In many real networks, degree distributions have a "long tail". That is, there is a relatively (to a random graph) high probability that nodes of degree much greater than the mean degree exist in the graph. Cumulative degree distribution graphs are often created to study the long tails of degree distributions because simple histograms of degree are generally quite noisy. This is because the small number of observations of nodes with high degree makes characterization of degree distribution tails difficult.

It is customary in the network literature to determine whether the degree distribution of the network of interest follows a power law. In this case, the probability of a node occurring with degree k decreases as k increases proportional to $k^{-\alpha}$. Networks that follow this trend tend to have high-degree nodes (sometimes called "hubs") connected to lower-degree nodes. This can have interesting implications regarding the relative importance a network's nodes to overall connectivity for both technological and social networks.

None of the degree distributions of the representations of the PDS networks appear to follow a power law. As some of the literature suggests, this is expected for affiliation networks because once an actor dies, he or she stops accruing links. This limits the extent to which preferential attachment can occur (Amaral et al, 2000). While this might explain the lack of power laws for the author representations, it does not necessarily do so for the other representations.

The α coefficients reported in Table 1 were extracted from a fit to the left portion of the degree distributions before the severe drop-offs (Figure 5). One of the degree distributions, the "technological" network of instrument hosts connected by authors, does appear to be exponential, as the cumulative plot on a lognormal scale is a straight line.

Rather than forcing the PDS degree distributions into the "power law" box, it is more useful to note their anomalous characteristics. Figure 6 shows the histograms of degrees for two of the network representations. The interesting feature of the degree distributions of the network is the existence of a *large* number of nodes with high degree. This occurs because of projects that involve collaboration of many authors on a single dataset or when one author works on many different datasets, depending on the

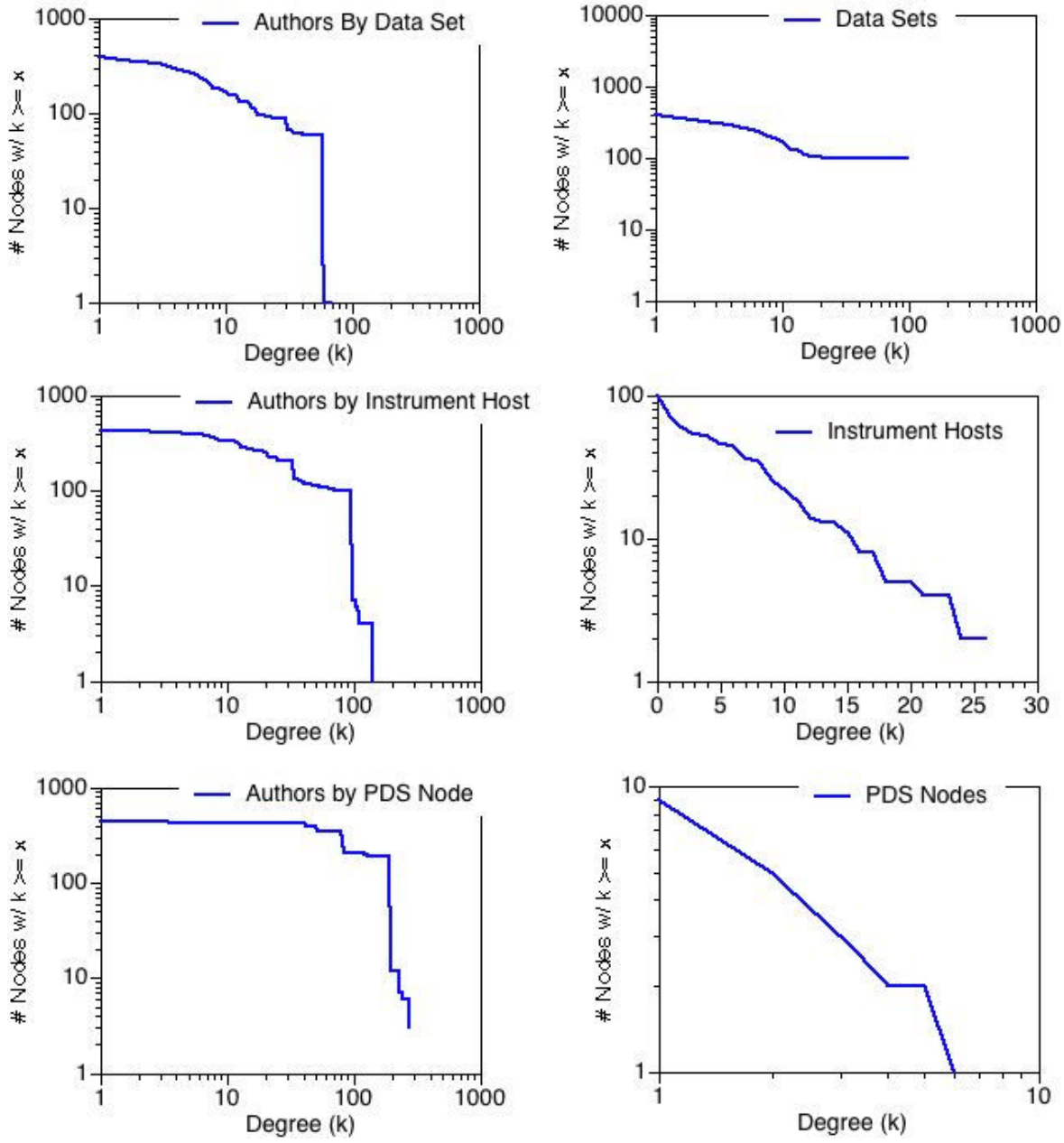

Figure 5. Cumulative degree distributions for the six unweighted PDS network representations. All except the middle-right graph of the instrument-host "technological" network are shown on a log-log scale. None clearly follow a power law, though the instrument-host-as-nodes network does follow an exponential distribution.

representation. Each of these authors, in an author-as-node representation, is symmetrically connected to all the other authors. This creates a symmetric cluster of nodes all with high degree. In the upper graph of Figure 6, there are 57 nodes with degree 57 because of a dataset involving more than 60 authors. Other authors of this dataset have degree higher than 60, and this connects that cluster to the remainder of the network. As discussed further below, this is also the reason that high degree correlations are observed (Table I) for this network in many of its representations.

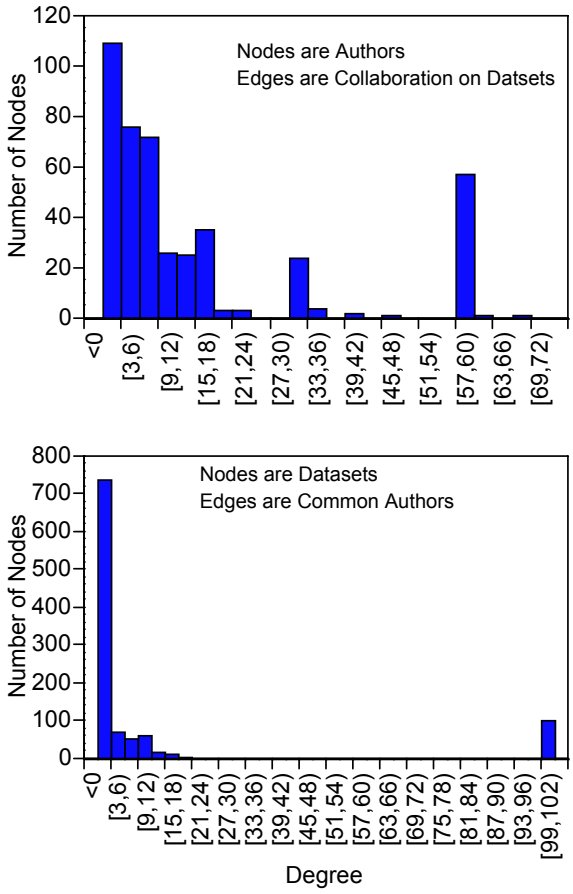

Figure 6. Histograms of degree for two PDS network representations.

3.4. Degree Correlation

3.4.1. Sign of the Degree Correlation

In the literature on networks, an interesting pattern has been observed in degree correlations of different types of networks. In particular, it seems that social networks tend to be assortative, meaning that they have positive degree correlations. Technological and biological networks tend to be disassortative (have negative degree correlations), and the assortativity of information networks is still more difficult to characterize. In his review paper, Mark Newman (2003) notes, "It is not clear what the explanation for this result is, or even if there is any one single explanation. (Probably there is not.)" Nevertheless, it still is interesting to consider the extent to which the networks derived from the PDS database follow these trends.

As stated previously, by most broad definitions, all of the networks with authors as nodes are social networks. For the purposes of this analysis, we identify the network of datasets as nodes to be an information network in which authors are the channels of

information between nodes, and we label the networks of PDS nodes and instrument hosts as technological networks by using the admittedly tenuous representation of authors as interconnections between engineered systems. According to these representations, we see positive degree correlations for all social and information networks regardless of weighting. For technological networks, the results are somewhat more mixed. For PDS nodes, the degree correlation is negative with and without weighting. For instrument hosts, the correlation is close to zero in both cases, though it is slightly positive for the unweighted network and slightly negative (with nearly identical magnitude) for the weighted network. In fact, weighting does not seem to have any appreciable effect on degree correlation since the magnitudes also seem to be similar with and without weighting in most of the cases.

As Newman states, there may not be any real meaning behind the sign of the degree correlation. Nevertheless, since we compared social, information, and technological representations of essentially the same network and saw some relation between network type and degree correlation, it may be that the sign of the correlation is dependent on the choice of network representation. While this result would agree with Newman's belief that there is no real meaning, it could be an indication of interesting differences between representations of a network. Still, there is other reason to question the validity of the degree correlation, which is discussed in the next subsection.

3.4.2. Limitations of the Degree Correlation

The network literature does not stress the limitations of using the summary statistic of degree correlation to describe networks. Correlation, or the Pearson correlation coefficient (r in Table 1) more specifically, is a method to measure the strength of the linear relationship between two variables. In this case, these two variables are the degree on either side of each edge in a network. One of the most serious, and well-known, limitations of the correlation coefficient is its sensitivity to extreme values. This is the problem that leads to the high values (close to 1) of r reported in Table 1.

A figure from an introductory statistics textbook best illustrates this problem (Chambers et al 1983). Figure 7 shows eight different scatter plots. The slope of the least squares regression line fit to the data is the Pearson correlation coefficient (r). In each of

these very different plots r is the same, 0.7. This is troublesome because the different patterns in these data could presumably represent quite different relationships between the two variables in question, and just reporting an r of 0.7 is misleading and incomplete. The Graph (6) of Figure 7 is the case in which the degree correlation is most meaningful: when the two variables are actually linearly associated with each other.

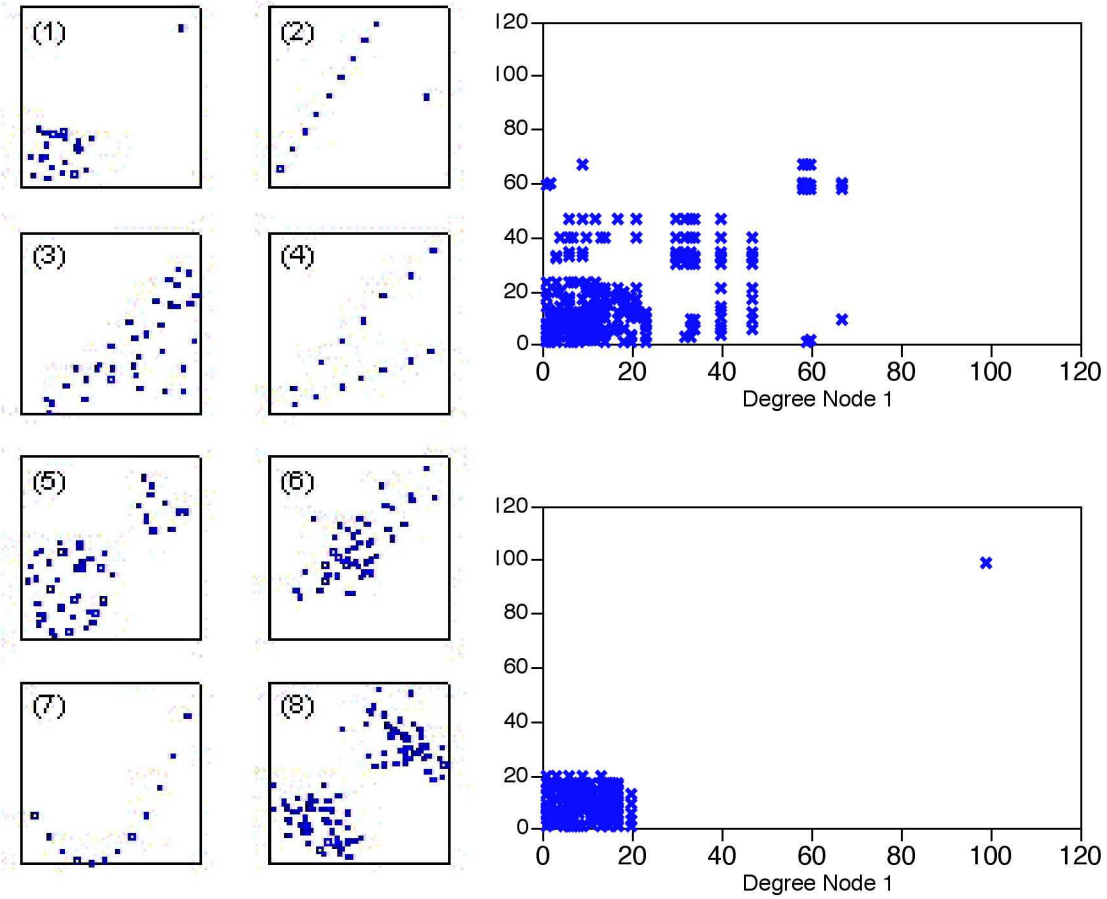

Figure 7. The eight small plots on the left represent datasets of two variables that all have correlation coefficients of 0.7. (*Source:* Chambers et al 1983) The two plots on the right are scatterplots of degree for adjacent edges in two representations of the PDS network. The upper graph is authors connected by collaboration on datasets (r = 0.96) and the lower graph is datasets connected by authors (r = 0.99). The extreme values (like the 99 points at degree 99 in the datasets graph) drive the correlation coefficient to close to one.

Figure 7 also shows the scatter plots of degree for the network of authors connected by collaboration on a dataset and for datasets connected by authors (the same networks shown in Figure 6). The reason the degree correlation is close to 1 in these cases is that the extreme values, the nodes with degree much larger than average,

influence the results considerably. As the histograms in Figure 6 show, the extreme values are not just one point but many. If the nodes with degree 99 are removed in the datasets network, for example, the degree correlation becomes about 0.5. If all but one of these points are removed, the degree correlation is about 0.7. It does not seem that removing these nodes from the calculation of degree correlation would be correct, however, because the nodes are a real, unique, and important part of the network.

4. Community Structure

Considering the structure of the PDS presented in Figure 1 and the emphasis placed on the PDS nodes, one might expect the collaborations of scientists on datasets submitted to the PDS to follow a pattern that maps to PDS nodes. For example, a hypothesis is that scientists tend to collaborate on datasets that "belong" to the same PDS node. This would mean that tightly knit areas in the affiliation network of authors connected by collaboration on datasets would map onto the PDS nodes (following Girvan and Newman 2002). Additionally, authors that contributed to datasets in multiple PDS nodes would be important in connecting the entire network. An alternative hypothesis is that the community structure (tightly knit areas of the author collaboration network) maps onto the instrument hosts. That is, authors tend to collaborate on datasets collected by the same spacecraft or telescope and that authors that work with more than one instrument host are key in connecting the network.

To test these hypotheses, we ran the Newman-Girvan (Girvan and Newman 2002) algorithm to find community structure in the author collaboration network for PDS (authors connected by collaboration on datasets). We found the second hypothesis to be true: the tightly knit communities correspond to different instrument hosts, and they do *not* correspond well to PDS node. Figure 8 shows the results of this test for the two largest connected components of this network. The node colors represent the instrument host and the shapes represent the results of the Newman-Girvan algorithm. The algorithm correctly places all but a few nodes in groups that correspond to instrument host.

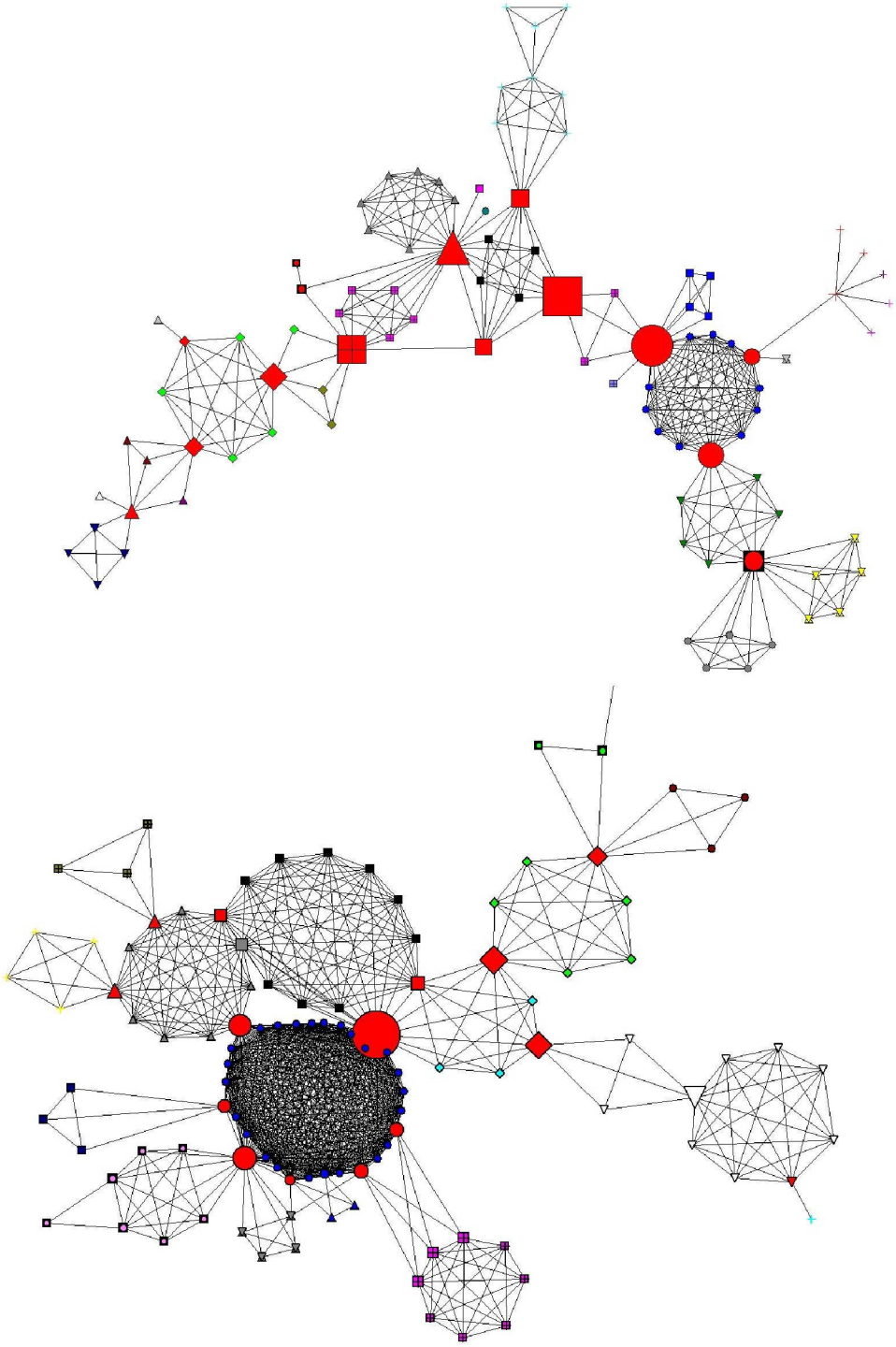

Figure 8. Results of the Newman-Girvan community structure algorithm (node shape) compared to instrument host affiliation (node color). Betweenness is represented by node size, and authors that worked on multiple instrument hosts are red nodes. This indicates that nodes with high betweenness are generally between clusters of authors that work on different instrument hosts, which is the reason Newman-Girvan algorithm works.

The finding that the Newman-Girvan algorithm "works" for instrument hosts but not for PDS nodes yields insight into how the authors of datasets actually work together. A missing piece of the puzzle is how the authors of papers that *use* the data collaborate. Unfortunately the citing mechanisms for the PDS are too inconsistent for a Web of Science search to be used to create a citation network around these publications. The "NASA Planetary Data System" is a listed citation in the Web of Science but it only returns two papers. The PDS administrators have only recently begun to stress a systematic citation procedure, and only two of the nodes have adopted it so far.

Despite this limitation, it is interesting that such an anonymous technique like the Newman-Girvan algorithm correctly identifies the community structure in the network of PDS dataset authors. It is intuitive that communities for dataset creation would surround the instrument hosts that are used to actually collect the data. It is also intuitive that the authors with largest betweenness (shown as node size in Figure 8) would be those that work with multiple instrument hosts (shown as red in Figure 8). It is an encouraging result for the Newman-Girvan algorithm that it correctly categorizes these communities. Girvan and Newman (2002) present similar tests of how well their algorithm works to identify "subject matter" communities (e.g. physics, economics) for other author affiliation networks like that from the citations of papers written by people affiliated with the Sante Fe Institute.

5. Centrality

One of the themes of the literature on networks, particularly social networks, is centrality. The best known social network in which centrality is a key issue is that of film actors. The common belief is that Kevin Bacon is somehow the center of the Hollywood universe. In reality, virtually any actor can appear to be the center of that network because it satisfies the small world conditions. The true center of the film actor network is Rod Steiger; Kevin Bacon ranks only 1,049.[4] The meaning of these results is limited. First, the only centrality metric used to determine the "best centers" in this network is closeness. Secondly, the weighting of the edges is not taken into account in the literature

[4] Tjaden, B. and Department of Computer Science, University of Virginia, *The Oracle of Bacon at Virginia*, <http://www.cs.virginia.edu/oracle/>, accessed April 28, 2006.

on this network.[5] Finally, the literature examines only the network of actors and does not look at the complementary network of movies – the most central movie might provide some insight about the network as well. In this section, we examine the centrality of the PDS network with special consideration to these aspects that are largely ignored in the literature.

5.1. Overall Centralization and Network Representation

The purpose of this subsection is to determine the relationship between network representation and overall centralization of the network, at least in the special case of the PDS network. The effects of the choice of 1-mode projection and of weighting both depend on the measure of centrality used. Table 2 summarizes the degree, closeness, betweenness, and eigenvector centrality of the 12 representations of the PDS network.

Table 2. Overall centrality measures for the 12 networks derived from the PDS database. Network representation and weighting both can affect some of the measures, though some of the measures are not valid for some of the networks.

			Centrality Measure			
Type of Node	Network	Weighted?	Degree	Closeness	Betweenness	Eigenvector
Authors as Nodes	PDS Nodes	No	0.339	0.364	0.136	0.060
	PDS Nodes	Yes	0.172	0.364	0.136	0.062
	Instrument Hosts	No	0.237	Unconnected	0.147	0.145
	Instrument Hosts	Yes	0.043	Unconnected	0.147	0.276
	Data Sets	No	0.120	Unconnected	0.026	0.354
	Data Sets	Yes	0.026	Unconnected	0.026	0.014
Events as Nodes	PDS Nodes	No	0.571	0.578	0.605	0.563
	PDS Nodes	Yes	0.314	0.578	0.605	2.152
	Instrument Hosts	No	0.205	Unconnected	0.074	0.385
	Instrument Hosts	Yes	0.046	Unconnected	0.074	1.555
	Data Sets	No	0.084	Unconnected	0.004	0.132
	Data Sets	Yes	0.006	Unconnected	0.004	0.512

The first interesting result enumerated in this table is that closeness centrality is an invalid metric for unconnected networks (in this case, all but the PDS networks). The reason for this is that the path lengths between nodes in disconnected components are infinite. However, this is exactly the metric that is used in the analysis of the film actor network. Because of this characteristic, the literature on film actors concentrates only on

[5] Of course, because weighting does not affect path length, it does not affect closeness centrality either. Nevertheless, weighting does affect other centrality metrics, as will be explained.

the giant connected component, which accounts for about 90 percent of the actors (Watts, 1999). Although this metric might still be useful given such a large main component, this is not necessarily the case for all social networks. Marchiori and Latora (2000) attempt to include the entire network in this type of analysis by incorporating their connectivity length metric, but the usefulness of this metric has not been tested in other works. Therefore, it is important that other centrality measures be used to characterize these networks.

As was discussed earlier in the paper, the weighing of edges in a network does not affect path length. As a result, closeness and betweenness centrality are the same for each network regardless of weighting. On the other hand, weighting does affect degree centrality. By thinking of a weighted edge as multiple edges (as explained previously), we can see easily that a weighted edge means that the two nodes to which it is connected have higher degree. The weighting of edges also affects eigenvector centrality, but it is not entirely clear that this metric is actually indicative of the structure of the network because it results in a centralization of greater than 100 percent for the weighted versions of the information network and both technological networks.

For both the weighted and the unweighted case, the most centralized of the 12 networks is the network of PDS nodes. The reason for this becomes clear with just a brief perusal through the web interface of the database.[6] Searching for datasets by "Curating Node" reveals that there are hundreds of datasets on a few nodes but only a handful on others. We will return to this aspect of the network when we discuss the most central nodes in the next subsection.

It appears from the results in Table 2 that weighting in these networks actually causes a decrease in overall degree centralization. The reason for this is apparently that the nodes with the highest degree are not the ones whose degree increases as a result of weighting. This result has some interesting implications. For example, consider any of the three representations of the author network. It appears that the scientists that work with the greatest number of other scientists are less likely to work with the same people multiple times. If the effect of weighting had been to increase overall centralization, the

[6] National Aeronautics and Space Administration, "Data Set Power Search," *Planetary Data System*, <http://starbrite.jpl.nasa.gov/pds/power.jsp>, accessed on March 6, 2006.

result simply would have reinforced the centralization of the network. Given the results that we found, though, it appears that ignoring the weighted edges in the network leads to an underestimate of the activity of nodes that interact frequently with each other (eg. authors that have worked together on multiple datasets).

Watts (1999) suggests that the small-world character of the network of film actors results, at least in part, from the existence of "linchpins" that span genres, eras, or countries. Similar linchpins working across planetary science disciplines probably play an important role in the PDS network (and in the planetary science community more generally) as well. Furthermore, on a purely statistical basis, those authors that work with a greater number of other authors could be more likely to fill this role. If this is the case, it actually might be appropriate to overvalue those nodes that are weakly connected to many nodes. On the other hand, it is possible that a linchpin author that works across disciplines collaborates only with one other author in each discipline and might do so on multiple occasions. Thus, there is unlikely to be a direct general relationship between weighting and the existence of linchpins, so it still is important to consider the effect of weighting on network centralization.

Still, although degree centrality decreases with weighting, the same does not seem to be true for eigenvector centrality. In fact, the apparent effect here is that the centrality actually increases for all of the networks except that of authors connected by datasets (though the increase is nearly negligible for the network of authors connected by PDS nodes). However, as discussed above, because the eigenvector centrality is greater than 100 percent for two of the networks, we are skeptical about the value of this metric. Given this and the lack of effect of weighting on closeness and betweenness centrality, we focus on degree centrality to comment on the overall effect of weighting on a network.

In an effort to quantify the effect of weighting on the structure of the network, we propose a new metric, W, weighting extent:

$$W = \log\left(\frac{C_{D,weighted}}{C_{D,unweighted}}\right), \tag{1}$$

where $C_{D,\,weighted}$ is the weighted degree centrality, and $C_{D,\,unweighted}$ is the unweighted degree centrality. In the case that $W = 0$, weighting has no effect on the centrality of the

network. Table 3 shows the weighting extent for each of the six representations of the PDS network. Note that $W < 0$ in all cases presented here.

Table 3. Weighting extent, *W*, for each of the six representations of the PDS network.

Type of Node	Network	Weighting Extent
Authors as Nodes	PDS Nodes	-0.296
	Instrument Hosts	-0.746
	Data Sets	-0.672
Events as Nodes	PDS Nodes	-0.260
	Instrument Hosts	-0.647
	Data Sets	-1.140

We propose this new metric with full awareness of its limitations. It implies a net total effect on the network based on just one measure of centrality. Therefore, the robustness of this metric will need to be tested against other networks. Furthermore, even with such testing, it is important (as with all network metrics) that the value of *W* be considered in the context of the specific network being analyzed. Still, this metric does capture the extent to which weighting affects the network, at least in terms of degree and centralization.

5.2. The "Best Centers" in the Planetary Data System

In this subsection, we present our analysis of the overall "best centers" in the Planetary Data System. The first step in this process was to calculate the most central actors in each of the 12 networks according to each of the four centrality measures. From each of these 48 calculations, we recorded those actors with the top two centrality scores.[7] This was an objective process based on the numerical results for each network.

Selecting the *overall* best centers from this long list of actors, however, was a bit more difficult. To do this, we simply looked through the resulting list for each network and picked the actors that most ranked among the most central according to more than one metric. In doing this, we gave less consideration to those measures that were likely to be invalid (e.g. closeness centrality for an unconnected network or eigenvector

[7] Note that this is distinct from selecting the top two. If there were 10 actors with the highest score and 10 with the second highest score, then a total of 20 actors were selected.

centrality for a weighted network). When a clear single best center emerged, as it did for the information network of datasets connected by authors, we chose that actor. If there was a close second (or third or fourth), we chose a few overall best centers but ranked them against each other according to the same procedure. Although it would be possible to derive a systematic algorithm to pick the best centers, the extent to which certain metrics are suppressed still would be somewhat subjective. Therefore, even with such an algorithm, the chosen centers are not necessarily the only possible centers. Nevertheless, the centers presented here are, at least within some small error, highly central to the entire PDS network.

5.2.1. Information and Technological Centers

The overall best center for the information network of datasets comes from the Mars Exploration Rover (MER) mission. This data set, labeled MER1-M-MI-5-MOSAIC-OPS-V1.0 in the PDS database, consists of microscopic imager mosaic images. MER1 refers to the instrument host, in this case one of the Mars Exploration Rovers, Opportunity, which was sent to a location on Mars known as Meridiani Planum. (MER2 refers to the other rover, Spirit.)[8] It is not surprising that the most central data set would be one containing photographic images of one of the most popular planetary missions currently in progress. Similarly, the most central instrument host in the network is one of the Voyager spacecraft. The identical Voyager 1 and 2 spacecraft together are the longest-running spacecraft mission in the history of planetary science and, so, have sent an enormous amount of data back to Earth. The second and third best centers in this network are simply two catch-all instrument hosts that refer to "Various Ground-based Telescopes" and "Public Literature."

[8] National Aeronautics and Space Administration, "Host Information," *Planetary Data System*, <http://starbrite.jpl.nasa.gov/pds/viewHostProfile.jsp?INSTRUMENT_HOST_ID=MER1>, accessed on May 14, 2006.

Table 4. Overall best centers for the information and technological representations of the PDS network.

Overall Best Centers - Information and Technological		
PDS Nodes	**Instrument Hosts**	**Data Sets**
Small_Bodies	Voyager 2	MER1-M-MI-5-MOSAIC-OPS-V1.0
Planetary_Atmospheres	Ground-Based Telescopes	
	Public Literature	

The best centers of the "technological" PDS node network is especially interesting because it provides a useful graphical representation, Figure 9, for comparison with the notional PDS architecture shown in Figure 1. According to the notional

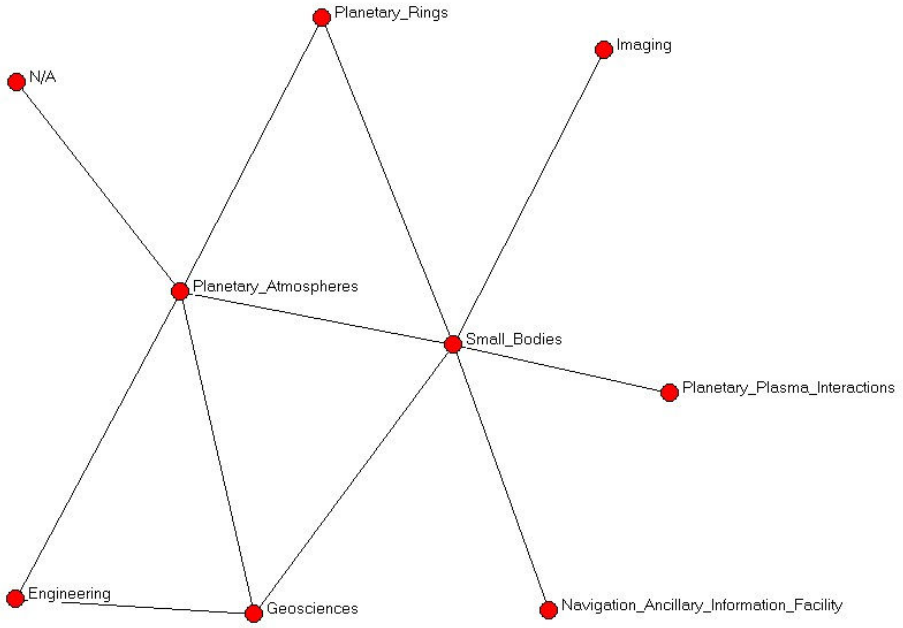

Figure 9. The network of PDS nodes connected by authors as edges. The two best centers of this network are Small Bodies and Planetary Atmospheres, respectively.

architecture, the network is a star shape, and all of the nodes are equally important. In the network in Figure 9, however, it is clear that Small Bodies and Planetary Atmospheres form the core of the system in terms of usage for the uploading of datasets, and the other nodes are largely peripheral. When we contacted scientists involved in the PDS before we began our data collection and analysis, we found that the Small Bodies Node was the most organized and able to provide us with information. Therefore, it was not a surprise that this turned out to be the most central of the nodes in the network.

5.2.2. What is Your Szego Number?

Unlike the networks of events as nodes, the "social" networks all have the same type of node – the authors. Therefore, it is possible to select overall best centers not just for each of the three social network representations but also for the entire PDS network. The results that we obtained, summarized in Table 5, indicate that Karoly Szego of KFKI, the home of the Hungarian Academy of Sciences, is the Erdös of the Planetary Data System. (It is interesting to note that Szego and Erdös are compatriots. Is it common for Hungarians to be at the center of scientific coauthorship networks?) Szego seems to be involved in the broader space community, as he has served on the committee of the International Conference on Low-Cost Planetary Missions held by the International Academy of Astronautics (IAA).[9] He is also the editor of a volume entitled, *The Environmental Model of Mars*, which contains 22 essays from the proceedings of the second Colloquium of the Committee on Space Research (COSPAR).[10]

Table 5. Overall best centers for the social representations of the PDS network. The last column lists the two authors that appear most in the centrality measures.

Overall Best Centers - Authors			
By PDS Nodes	**By Instrument Hosts**	**By Data Sets**	**Overall**
T. Z. Martin	C. Neese	K. Szego	K. Szego
R. Mehlman	L. S. Elson	J. T. Gosling	J. T. Gosling
J. R. Spencer	C. H. Acton	R. F. Beebe	
	B.V.Semenov		

Apparently, though, Szego does not have any particular role in the management of the PDS, which makes sense since he is not American. However, some of the other top centers listed in Table 5 are actively involved. For example, Carol Neese, one of our primary PDS contacts, is the coordinator of the Asteroids subnode within the Small Bodies Node. Reta Beebe, another of our contacts, appears twice on the PDS organization chart, which is shown in Figure 10. In addition, as can be seen in the organization chart, Charles Acton is the manager of the NAIF node. Still, none of the

[9] "First Announcement: Fifth IAA International Conference on Low-Cost Planetary Missions," *euSpaceRef.com*, <http://eu.spaceref.com/news/viewpr.html?pid=9358>, accessed on May 15, 2006.
[10] "The Environmental Model of Mars," *Elsevier*, <http://www.elsevier.com/wps/find/bookdescription.librarians/28785/description#description>, accessed on May 15, 2006.

three best centers in the network of authors connected by PDS node actually appear in the organization chart. Nevertheless, there does seem to be some (though admittedly imperfect) relationship between centrality in the network and management of the PDS.

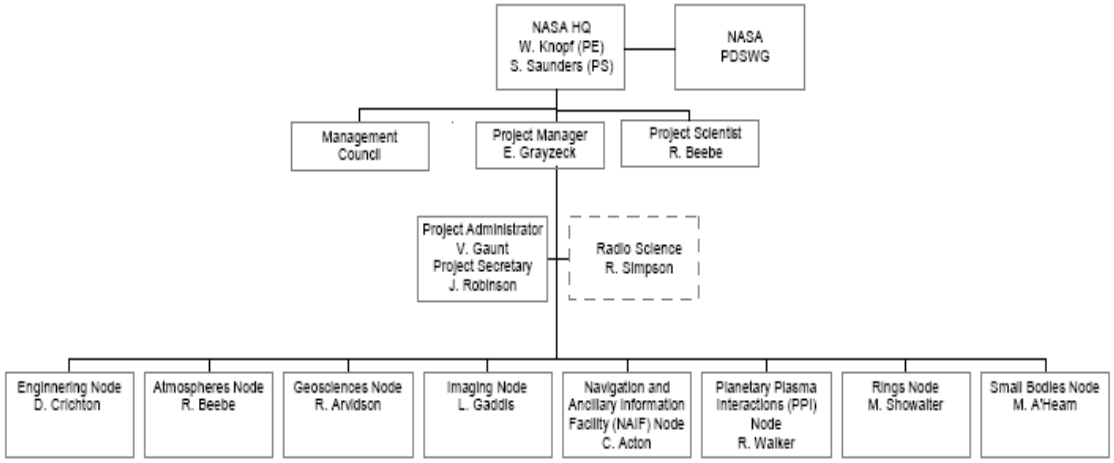

Figure 10. Organization chart of the Planetary Data System.
Figure courtesy of NASA. *Source:* National Aeronautics and Space Administration, *Planetary Data System (PDS) Project Organization*, <http://pds.jpl.nasa.gov/tools/pds_org.pdf>, accessed on February 28, 2006.

6. Conclusions

6.1. General Observations…

6.1.1. …About the Planetary Data System

In this study, we have analyzed the collaboration patterns of scientists that help to generate and upload data onto the Planetary Data System. We found that these scientists form tight-knit communities around instrument hosts rather than around the notional PDS nodes. In addition, the measures of centrality, combined with our own experiences interacting with the node managers, suggest that the Small Bodies Node is well-run and important in connecting the PDS nodes together. This suggests that scientists that study small bodies (ie. asteroids, comets, and dust) are likely to participate in the provision of data to other PDS nodes as well. Still, it is difficult to draw robust and generalized conclusions about the system because our analysis includes only uploaded data (without incorporating data access patterns) at a snapshot in time. To make a meaningful statement about the value of the system to the scientific community and to the general public, these aspects of the system would have to be considered as well. Still, this project provides the basis for further study of the PDS.

Furthermore, it is important to note that the Planetary Data System represents only a small cross-section of the entire planetary science community. Although NASA is attempting to archive all planetary data on this system, a great deal of work still needs to be done. Moreover, many of the datasets are in the system are not appropriately cited, which makes it difficult or impossible to know who the authors are. Also, there are, of course, a great deal more than 439 planetary scientists in the world. Nevertheless, data do exist on broader segments of the planetary science community, and we will return to this yet-unexploited opportunity in subsection on future work.

6.1.2. ...About the Science of Networks

This study's primary conclusions about the analysis of networks relate to the shortcomings in the metrics currently used to compare different networks at a general level. Many of these limitations arise from the need to discard important information to make simple and broad comparisons between networks. Examples of this include the weighting of edges in affiliation networks according to the number (or strength) of interactions and the use of the Pearson degree correlation without accompanying scatter-plots.

The weighting of edges is important because it carries information about the frequency and strength of the connections between actors or events in the network. This weighting, however, is often ignored in an attempt to capture the essence of social interactions in a generalizable way. This is done for good reason – many of the commonly used metrics become essentially meaningless when weighting is included. With the result that weighting can decrease centralization, at least in the PDS network, we suggest a metric to assess the extent to which weighting affects degree and centralization.

The degree correlations of many of the network representations of the PDS network are close to 1. This occurs because of the existence of large symmetric clusters in which all nodes having higher than average degree. This result suggests that the Pearson correlation coefficient, r, can produced misleading results about the nature of the network. Therefore, this metric should be used cautiously and only in conjunction with degree-versus-degree scatter-plots to avoid reporting meaningless values of r.

The results of our community structure analysis support the use of the Newman-Girvan algorithm to find tightly knit communities within affiliation networks. The algorithm's results for the social network of authors connected by datasets matched well to the communities defined by the instrument hosts on which the authors worked. Nodes with high betweenness correspond almost exactly to authors that worked on multiple instrument hosts.

Finally, our results indicate that the choice of network representation affects the magnitudes of most of the common metrics used to evaluate networks. To mitigate the effect of weighting, we suggest the use of certain metrics that incorporate this phenomenon. Examples include Marchiori and Latora's connectivity length and the weighting extent proposed in this paper. Even with weighting taken into account, though, the choice of which system components map to nodes and which to edges still affects the results of the network analysis. This stresses the importance of maintaining network context – one should think carefully about the questions of interest and try to choose the appropriate network representation.

6.2. Future Work

6.2.1. *Analysis of a Subject Area Network, including Network Growth*

One possibility for future work is to study the sub-network within one of the PDS nodes to determine centrality and community structure within a subfield. Another interesting aspect of such a study would be to determine if these sub-networks are more or less connected than the overall network. In addition, with narrower scope of the network, it might be more straightforward to examine the evolution and growth of the network over time. This could be done by incorporating the dates on which datasets were uploaded into the system (the PDS reports release dates for each of the datasets).

6.2.2. PDS Data Access Statistics

During our initial research, we found that the PDS node managers report download usage statistics to the central PDS office. We requested a copy of these statistics from the central office, but we had trouble actually getting the data. Some of the people that we contacted told us that the statistics did not exist (although we already

had the data for the Small Bodies Node – yet another indication of the coordination within that node). Others referred us to the individual node managers, who, in turn, sent us back to the central office. Nevertheless, we eventually were able to obtain this information for four of the nodes. The problem with these statistics, though, is that NASA collects them only at the PDS-node level (not for individual datasets or instrument hosts), and the only information available about each user is the hostname of the computer used to access the data. Although it might be possible, at least technically, to collect more detailed information about the data being accessed, the identities of the scientists necessarily cannot be known. Still, some limited insight might be made from the access of PDS nodes by various institutions.

Another approach to analyzing the usage of PDS data would be to examine papers published using these data. The PDS does provide scientists with a standard citation format for these data. The format, however, has been in use only since 2003, and even since then, it has not been used consistently. Still, a search through Web of Science might provide some papers whose collaboration network could be analyzed.

6.2.3. Social Network Analysis of a Major Planetary Science Conference

Richard P. Binzel, one of MIT's own nodes in the PDS collaboration network, has suggested that it might be useful and interesting to conduct a similar study using the papers submitted to a major planetary science conference. As discussed previously, the PDS provides only a limited portion of the entire planetary science community. According to Professor Binzel, however, the Annual Meeting of the American Astronomical Society's (AAS's) Division for Planetary Sciences (DPS) would provide a much larger and more representative sample. Furthermore, he believes that the subject areas of the PDS nodes map well to the disciplines around which DPS meetings are organized.[11] Therefore, it likely would be feasible not only to repeat the analysis for the DPS data but also to compare results to those obtained for the PDS. This analysis could include an assessment of whether the community structure and centrality discussed in this paper are general phenomena or just unique to the PDS.

[11] Richard P. Binzel, personal communication, May 6, 2006.

6.2.4. Funding and Political Support of Planetary Spacecraft

Our last and least-defined suggestion for future work is to gather data of some sort on the funding and political support for missions and/or for the subject areas of the PDS nodes. The intent would be to determine the dynamics between scientists, engineers, and policymakers in the lifecycle of planetary missions. The intent would be that the results would relate somehow to the conclustions presented in this paper. Such a study might help to understand the origins of political support for planetary missions and for the PDS. The data would be difficult to obtain, but the results of such a study could have the potential to break new ground in science policy and space policy in the United States.

References

Amaral, L.A.N., Scala, A., Barthélémy, M., and Stanley, H.E., "Classes of small-world networks," *Proc. Natl. Acad. Sci. USA* **97**, 11149-11152 (2000).

Borgatti, Everett, & Freeman, **UCINet** 6 Network Analysis Software. Analytic Technologies, 11 Ohlin Ln., Harvard, MA 01451 (2002).

Chambers, Cleveland, Kleiner, and Tukey, *Graphical Methods for Data Analysis,* Duxbury Press (1983).

Erdös, P. and A. Renyi, "On random graphs, I" *Publicationes Mathematicae* **6** (1959), 290-297.

Girvan, M. and M.E.J. Newman, "Community structure in social and biological networks," Proc. Natl. Acad. Sci. USA 99, 7821-7826 (2002)

Marchiori, M. and Latora, V., "Harmony in the small-world," *Physica A* **285**, 539-546 (2000).

Newman, M.E.J., "The structure and function of complex networks," *SIAM Review* **45**, 167-256 (2003).

Newman, M.E.J., Strogatz, S.H., and Watts, D.J., Random graphs with arbitrary degree distributions and their applications, *Phys. Rev. E* **64**, 026118 (2001).

Watts, D. J. (1999) *Small Worlds: The Dynamics of Networks Between Order and Randomness* (Princeton Univ. Press, Princeton, NJ).

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

Appendix

Table A-1. The "best centers" using all four metrics for each of the 12 representations of the PDS network. Each type of node is represented by a unique identifier assigned to it for the purposes of this study. The identifier lookup tables are available from the authors.

			Top Two Best Centers											
			Degree			Closeness			Betweenness			Eigenvector		
Type of Node	Network	Weighting	Node ID	Degree Value	Normalized	Node ID	Farness Value	Normalized	Node ID	Value	Normalized	Node ID	Value	Normalized
Authors as Nodes	PDS Nodes	Unweighted	65, 153, 74	273	0.623	65, 153, 74	603	0.726	7	13197.22	0.138	74, 153, 65	0.073	0.103
			36, 76, 134	271	0.619	36, 76, 134	608	0.720	76, 36, 134	9308	0.097	36, 134, 76	0.072	0.103
	PDS Nodes	Weighted	65, 153, 74	275	0.314	65, 153, 74	603	0.726	7	13197.22	0.138	84, 90, 91, 130, 92	0.074	0.105
			36, 76, 134	273	0.312	36, 76, 134	608	0.720	76, 36, 134	9308	0.097	65, 153, 74	0.074	0.104
	Instrument Hosts	Unweighted	2	141	0.322	150	12216	0.036	2	14359.02	0.150	84	0.076	0.108
			32, 10, 9	140	0.320	9, 10, 32	12226	0.036	3	10044.7	0.105	24	0.057	0.081
	Instrument Hosts	Weighted	2	169	0.055	150	12216	0.036	2	14359.02	0.150	33	0.204	0.288
			32, 10, 9	160	0.052	9, 10, 32	12226	0.036	3	10044.7	0.105	24	0.162	0.229
	Data Sets	Unweighted	76	67	0.153	61	144631	0.303	61	2504.33	0.026	85	0.251	0.354
			134	60	0.137	106	144654	0.303	36	2379	0.025	41	0.121	0.171
	Data Sets	Weighted	11, 18, 21, 26, 20, 15, 19, 14, 13, 23, 17, 12, 27, 25, 28, 6	120	0.030	61	144631	0.303	61	2504.33	0.026	--	All <= 0	All <= 0
			76	76	0.019	106	144654	0.303	36	2379	0.025	--	All <= 0	All <= 0
Events as Nodes	PDS Nodes	Unweighted	Small_Bodies	6	75	Small_Bodies	10	0.8	Small_Bodies	18.5	0.661	Small_Bodies	0.529	0.748
			Planetary_Atmospheres	5	62.5	Planetary_Atmospheres	11	0.727	Planetary_Atmospheres	10.5	0.375	Planetary_Atmospheres	0.519	0.733
	PDS Nodes	Weighted	Small_Bodies	16	0.4	Small_Bodies	10	0.8	Small_Bodies	18.5	0.661	Small_Bodies	0.999	1.412
			Planetary_Atmospheres	14	0.35	Planetary_Atmospheres	11	0.727	Planetary_Atmospheres	10.5	0.375	Planetary_Rings	0.034	0.048
	Instrument Hosts	Unweighted	H98, H94	26	0.255	H98	3919	0.026	H98	405.86	0.079	H98	0.299	0.423
			H83, H31	23	0.225	H94	3927	0.026	H83	370.80	0.072	H94	0.275	0.389
	Instrument Hosts	Weighted	H94	61	0.054	H98	3919	0.026	H98	405.86	0.079	H9	0.994	1.405
			H98	45	0.040	H94	3927	0.026	H83	370.80	0.072	H44	0.052	0.074
	Data Sets	Unweighted	***	99	0.095	***	989615	0.001	ID0078	2222	0.004	***	0.1	0.141
			ID0545	20	0.019	ID0638	996066	0.001	ID0638	2134	0.004	All Others	0	0
	Data Sets	Weighted	ID0051-ID0053, ID0026-ID0028, ID0049, ID0050	115	0.007	***	989615	0.001	ID0078	2222	0.004	ID0026-ID0028, ID0049-ID0053	0.353	0.500
			***	99	0.006	ID0638	996066	0.001	ID0638	2134	0.004	ID0036-ID0038	0.023	0.032

*** = ID0384-ID0508 except ID0400-ID0404, ID0410-ID0413, ID0429, ID0444-ID0448, ID0465-ID0469, ID0475-ID0478, ID0494 Red indicates invalid or questionable results.

Table A-2. Selected sample of entries from which the node lists for this project were derived.

Data Set ID	Data Set Name	Version	Data Set Description / Long Name	InstrumentHost	Node	Subnode	Data Set ID	Author1	Author2	Author3	Author4	Author5
SAKIG-C-IMF-3-RDR-HALLEY-V1.0	SAKIGAKE INTERPLANETARY MAGNETIC FIELD DATA V 1.0	1.0	Sakigake satellite magnetic field	SAKIG	SBN	COMET	SAKIG-C-IMF-3-RDR-HALLEY	K.-I.Oyama				
SAKIG-C-SOW-3-RDR-HALLEY-V1.0	SAKIGAKE SOLAR WIND EXPERIMENT DATA V1.0	1.0	Sakigake satellite solar wind data	SAKIG	SBN	COMET	SAKIG-C-SOW-3-RDR-HALLEY	K.-I.Oyama				
SDU-A-NAVCAM-2-EDR-ANNEFRANK-'	STARDUST NAVCAM IMAGES OF ANNEFRANK	1.0	The Annefrank data set is a colle	SDU	SBN	ASTEROID	SDU-A-NAVCAM-2-EDR-ANNEFRANK	R.L.NewburnJr.				
SDU-C/D-CIDA-1-EDF/HK-V1.0	STARDUST CIDA DATA	1.0	Collection of time-of-flight spect	SDU	SBN	COMET	SDU-C/D-CIDA-1-EDF/HK	J.Ryno	B.V.Semenov	J.Kissel	J.Silen	C.H.Acton
SDU-C-DFMI-2-EDR-WILD2-V1.0	STARDUST DFMI WILD 2 ENCOUNTER EDR DATA	1.0	EDR data collected by the STARI	SDU	SBN	COMET	SDU-C-DFMI-2-EDR-WILD2	B.V.Semenov	A.J.Tuzzolino	J.A.M.McDonnell	H.W.Taylor	C.H.Acton
SDU-C-NAVCAM-5-WILD2-SHAPE-MC	TRI-AXIAL ELLIPSOID MODEL OF COMET WILD 2	1.0	Basic tri-axial ellipsoid shape moc	SDU	SBN	COMET	SDU-C-NAVCAM-5-WILD2-SHAPE-MC	T.Duxbury	T.L.Farnham			
SDU-C-SPICE-6-V1.0	STARDUST SPICE KERNELS V1.0	1.0	Navigation and ancillary data in the for	SDU	NAIF	ASTEROID	SDU-C-SPICE-6	B.V.Semenov	L.S.Elson	C.H.Acton		
STARDUST-C/E/L-DFMI-2-EDR-V1.0	STARDUST C/E/L DUST FLUX MONITOR INSTRUMENT-2-EDR	1.0	Data collected by the Dust Flux I	SDU	SBN	COMET	STARDUST-C/E/L-DFMI-2-EDR	H.W.Taylor				
STARDUST-C/E/L-NC-2-EDR-V1.0	STARDUST NAVCAM EARLY CRUISE IMAGES	1.0	Early cruise images from the Star	SDU	SBN	CALIBRATI	STARDUST-C/E/L-NC-2-EDR	C.Hash				

Table A-3. Selected sample of datasets (in the first column) with authors listed horizontally. In this table, the unique identifiers assigned for this study have been used.

ID	Name1	Name2	Name3	Name4	Name5	Name6	Name7	Name8	Name9	Name10
ID0638	36	292	55	169	273	202	7	307	133	434
ID0235	182	72	73	135	106	294	205	206	274	257
ID0827	129	119	118	165	99	94	114	100	76	128
ID0837	129	119	118	165	99	94	114	100	76	128
ID0030	436	276	417	397	389	432	392	210		
ID0078	77	141	5	161	147	120	302	346		
ID0128	350	393	338	309	5	365	190	216		
ID0942	385	270	130	192	249	168	91	90		
ID1040	198	233	364	175	212	69	153	398		
ID0085	117	97	58	60	84	237	321			
ID0088	236	187	214	316	293	42	74			
ID0114	5	330	30	245	160	156	154			