Suthers, D. D., Fusco, J., Schank, P., Chu, K.-H., & Schlager, M. (2013). Discovery of community structures in a heterogeneous professional online network. *Proceedings of the Hawaii International Conference on the System Sciences (HICSS-46), January 7-10, 2013, Grand Wailea, Maui, Hawai'i (CD-ROM).* New Brunswick: Institute of Electrical and Electronics Engineers, Inc. (IEEE).

Discovery of Community Structures in a Heterogeneous Professional Online Network

Dan Suthers
University of Hawai`i at Manoa
suthers@hawaii.edu

Judi Fusco
SRI International
judith.fusco@sri.com

Patricia Schank
SRI International
patricia.schank@sri.com

Kar-Hai Chu
University of Southern California
karhai.chu@usc.edu

Mark Schlager
Samsung Information Systems America
m.schlager@sisa.samsung.com

Abstract

Socio-technical networks that are heterogeneous in composition of actors and the media through which they interact are becoming common, but opportunities to study the emergent community structure of such networks are rare. We report a study of an international online network of educators involved in many forms of professional development and peer support, including sponsored and volunteer-driven activities taking place in both synchronous and asynchronous media, with participants from diverse career stages and occupations in education. A modularity-partitioning algorithm was applied to a directed, weighted, multimodal graph that represents associations between actors and the artifacts (chats, discussions and files) through which they interact. This analysis simultaneously detects cohesive subgroups of actors and artifacts, providing rich information about how communities are technologically embedded. Researchers deeply familiar with the network validated the interpretability of the partitions as corresponding to known activities, while also identifying new findings. The paper describes this interpretative validation, summarizes findings concerning the distribution and nature of communities and groups found within the larger heterogeneous network, and discusses open research questions and implications for practitioners.

1. Introduction

Many aspects of our social, occupational, and business lives are becoming increasingly embedded in "online" settings via Internet and mobile technologies [6, 14, 18, 25]. These environments can be heterogeneous along many dimensions, including participant characteristics, their purposes in participating, the available media affordances for interacting, and organized versus organic activities. Although heterogeneous networks are becoming quite common, few opportunities exist to rigorously study them due to increasing data restrictions: in addition to privacy concerns, data is increasingly seen as a strategic asset not to be disclosed to outsiders such as researchers.

Practitioners and researchers have been creating and studying online networks for more than two decades [21, 18, 28] showing the rich networks of support and high quality of dialogue people can achieve online. We know that relationships develop online, but in a complex online environment, many different means of interacting are available (e.g., text chat, discussion boards, voice chat, video chat, blogging, and tagging and sharing media) and there are qualitative differences between these types of online interactions [11, 14]. Such accounts have shown what is attainable in particular contexts, but we are still unable to rigorously measure their value, much less predict or guide results reliably or at scale [19]. To realize the promise of online networks, we must employ a new generation of methods that bridge multiple research traditions and types of data, and apply these methods to give leaders and members of these networks more insight to what they are accomplishing online.

This paper reports findings from an ongoing analysis of one large-scale heterogeneous network and our attempts to answer these questions. For nearly two decades, co-authors of this paper have supported SRI International's Tapped In® (tappedin.org), an international online network of educators involved in diverse forms of informal and formal professional development and peer support [9, 20]. Development of Tapped In was motivated by the desire to understand how to initiate and manage large heterogeneous communities of educators, how they evolve, and the benefits that participants and sponsors derive from their involvement.

This network includes activities that are sponsored by formal organizations (e.g., universities, school

districts, and nonprofits) mixed with volunteer driven and other unsponsored activities, in both synchronous and asynchronous media, with participants from across all career stages and diverse occupations related to education. Thus it provides a valuable opportunity to develop and test hypotheses, tools, and techniques for understanding heterogeneous networks. Cumulatively, Tapped In has hosted the content and activities of more than 150,000 education professionals (over 20,000 per year in our study period) in thousands of user-created spaces that contain threaded discussions, shared files and URLs, text chats, an event calendar, and other tools to support collaborative work. Over its history, more than 50 organizations, including education agencies and institutions of higher education, have consulted with Tapped In staff and became "tenants" in the system to meet the needs of students and faculty with online courses, workshops, seminars, mentoring programs, and other collaborative activities. While these organizations typically set up private spaces for people affiliated with them, there were also approximately 40-60 public activities per month designed by Tapped In members and open to anyone in the community (including tenant members). Volunteers drive the majority of Tapped In community-wide activity [9]. Extensive data collection capabilities underlying the system captured the activity of all members and groups including chat data, discussion board interactions, and file sharing. We selected a period of peak usage that occurred from September 2005 through May 2007 for analysis in this study.

Because Tapped In is populated with members of multiple tenant organizations as well as unaffiliated members, it is best seen as a network of education professionals rather than a single "community." Members may move freely between most forms of participation. The question of what communities (or other types of groups) exist in this network is a matter for empirical investigation. We approach this question in terms of the artifact-mediated associations found between members. In the present study, associations between actors and the artifacts (chat rooms, discussion forums, and files) through which they interact were used to detect cohesive subgroups of actors and artifacts, providing richer information than approaches that operate on networks of actors alone. We examined sociometric properties of these groups in relation to their size and classifications as unsponsored or sponsored by network tenant organizations, to determine whether there are any systematic variations in how groups of different sizes or different levels of sponsorship operate. Then, researchers deeply familiar with the network validated the interpretability of the findings in terms of known community or group activities, while also identifying new findings.

Thus, the study addresses instrumental and primary research questions, respectively: Are clusters of actors and artifacts found through graph-theoretic methods interpretable by humans familiar with Tapped In? If so, what do we learn about the structure of this heterogeneous network by analyzing detected clusters? The remainder of the paper begins by describing the method, as it offers a unique combination of existing representations and algorithms for social network analysis. Then the distribution of partitions found within the network as a whole is reported with respect to media use and sociometrics, to begin to answer the question of what a heterogeneous socio-technical network looks like. The partitions are interpreted to simultaneously provide examples of the kinds of communities we found embedded within the Tapped In network and demonstrate that the method is valid in the sense of providing interpretable results. We conclude with implications for research and practice.

3. Analytic Approach

We begin by describing the representation we work with, followed by conversion of data sources into this representation and the algorithms applied to this representation to characterize potential community activity within the network.

3.1. Associograms

Socio-technical networks are commonly studied using the methods of social network analysis, using sociograms or sociomatrix representations of the presence or strength of ties between human actors, and graph algorithms that leverage the power of this representation to expose both local (ego-centric) and nonlocal (network) social structures [27]. Singular representations of a tie between two actors summarize yet obscure the many interactions between the actors on which the tie is based, as well as the media through which they interacted. Our method seeks to retain the advantages of graph computations on a summary representation while retaining some of this information about how the actors interacted.

To do so, we use *bipartite, multimodal, directed weighted graphs*, similar to but more specific than affiliation networks. They are bipartite because all edges go strictly between actors and artifacts; and multimodal because the artifact nodes can be categorized into different kinds of mediators that they represent, in our case including chat rooms, discussion forums and files. Directed edges (arcs) indicate read/write relations or their analogs: an arc goes from an actor to an artifact if the actor has read that artifact, and from an artifact to an actor if the actor modified

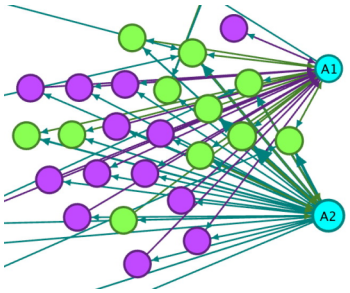

Figure 1. An Associogram. Actors represented by nodes on the right have read and written to the files and discussions represented by differently colored nodes on the left.

the artifact (the direction indicates a form of dependency, the reverse direction of information flow). Weights on the arcs indicate the number of events that took place between the corresponding actor/artifact pair in the indicated direction. Since "affiliation network" is not specific enough and "bipartite multimodal directed weighted graph" is too long, to highlight their unique nature we call these graphs "associograms" [24]. This term is inspired by Latour's [13] concept that social phenomena are "assembled" by dynamic networks of *associations* between human and non-human actors ("actants").

What is summarized as a tie in a sociogram is represented as a network of artifact-mediated associations in an associogram. For example, Figure 1 shows an actual portion of an associogram from Tapped In data, representing asymmetric interaction between two actors, with one actor writing most of the files and another writing to most of the discussions. A sociogram consisting of a single link between actors would fail to capture this information. The network directly retains information about the distribution of activity across media. Network analytic methods can then *simultaneously* tell us how both human actors and artifacts participate in generating the larger phenomena of interest, such as the presence of communities of actors and the media through which they are technologically embedded [14]. Although interaction is not directly represented, the associogram also provides a bridge to the interaction level of analysis [23]: it allows us to retrieve activity in specific media settings.

3.2. Data Preparation

Tapped In data was prepared as follows. We parsed and filtered databases and logs of user activity involving files, asynchronous threaded discussion forums, and synchronous chat rooms. Private chats were excluded from our analysis, which focuses on observable public behavior. All student activity in the K-12 (student) campus was excluded, as our research (and human subjects permission) focuses on the professional community. Guest accounts were also filtered, as different rotating individuals use these.

Associograms for each artifact type were constructed. A file node represents a single file, a discussion node an entire threaded discussion, and a chat node a chat room. A file node points to an actor if the actor created (uploaded) that file; an actor node points to a file node if the actor downloaded the file. A discussion node points to an actor node if the actor posted a message in the corresponding discussion forum; an actor node points to a discussion node if the actor loaded the discussion page. A chat room points to an actor if the actor posted a chat contribution in that chat room while someone else was present; an actor node points to a chat node if the actor was present in the room when another actor posted a chat contribution. Each of these arcs was weighted according to the number of times the events just described were seen. The data comprises 35,012 actants (represented by nodes in the graph), with 179,703 associations between them (represented by edges). There are 16,569,971 events included (each contributing a weight of 1 to some edge).

The associograms were exported in a format (VNA) readable by social network analysis software. Attributes for the different entities and weights on arcs were included in the export. Associograms for all artifact types were merged into a master associogram in Gephi 0.8.1 [2], where they were visualized using the OpenOrd algorithm [15] and partitioned as discussed in the next section.

3.3. Partitioning into Cohesive Subgroups

We use "community" or "group" to refer to empirically associated actants for whom it is also possible to identify some shared activity or purpose. This sense of "community" is much looser than the traditional *gemeinschaft* [26], and does not make claims about participant's own identities [7], but a more inclusive definition is appropriate for networked society [6, 29]. We use graph theoretic terms (e.g., "partition") when discussing algorithmic results that are candidates for interpretation as a particular kind of cohesive subgroup, and reserve "community" for when we are entering into the realm of such interpretation.

In the network analysis literature, "community detection" refers to finding subgraphs of mutually associated vertices under graph-theoretic definitions rather than to the sociological concept. A good graph-theoretic definition should capture the intuition that individuals in a sociological community are more closely associated with each other than they are with individuals outside of their community. Algorithms based on the modularity metric are widely used in the

literature for this purpose. The modularity metric compares the density of weighted links inside (non-overlapping) partitions of vertices to weighted links expected in a random graph, to find highly modular partitions. Finding the best possible partition under a modularity metric is computationally hard (impractical to compute on large networks), but Gephi includes a fast algorithm by Blondel et al. [3] that has been shown to give good approximations. We chose to use this algorithm because the modularity metric is well known, and a fast implementation that can handle our large network is available.

4. Characterizing the Network

4.1. Partitions Obtained

The modularity-partitioning algorithm identified 234 partitions with a modularity of 0.828 (the maximum is 1; this value indicates strong clustering). We analyzed three groups of these partitions. First, our complete analysis (sociometrics and human interpretation) was applied to the top 19 partitions, those that contain at least 1% of the actant nodes (ranging from 344 to 11252 nodes and 644 to 36949 edges). Together, the top 19 partitions contain about 75% of the nodes. Second, full sociometric statistics were also computed for all the remaining partitions of size at least 0.1% of the actants (86 partitions). Third, 21 partitions were sampled from the 215 partitions not in the top 19, by taking every 10th partition, and sociometrics and human interpretation were undertaken for all of these partitions. Sampling was done because we wanted to characterize the changes in partitions as they get smaller in the "long tail", but it was infeasible to conduct time-consuming human interpretations of all of the 234 partitions, and many were too small to be of interest. Sampling was done at regular intervals rather than randomly because we wanted to characterize the size distribution and associated changes in other variables. The sampled partitions each contain less than 1% of the actant nodes (ranging from 2 to 303 nodes and 2 to 2070 edges). In summary, sociometrics and human interpretation were applied to a total of 40 partitions, 19 from the top 1% and 21 from the bottom 99%, and sociometrics were applied to all of the top 86 partitions (down to 0.1% in size).

Figure 2 shows an OpenOrd visualization of the partitions obtained. Each partition is given a different color, and structures corresponding to major candidate communities are visible. Generally the visualization is not interpreted in this form, but rather filters are applied to view partitions independently of each other, and local structures are examined by zooming in to see how actors and artifacts are relating to each other.

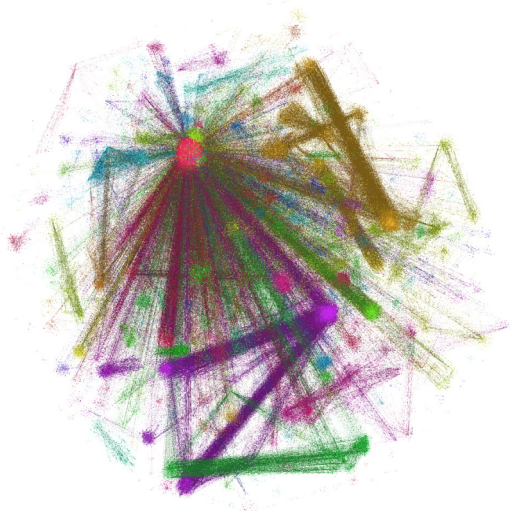

Figure 2. OpenOrd visualization of partitions found in combined associogram for actors associated via chats, discussions and files.

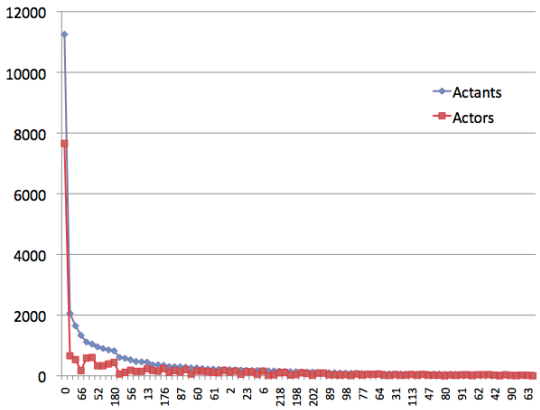

Figure 3. Distribution of size of partitions.

Table 1 shows the top 19 partitions found, with selected statistics (see table for legend; further explanation provided below).

4.2. Distribution of size of partitions

Graphing the size distribution of partitions reveals an exponential decay, with a large initial partition followed by decreasing sized partitions and a long tail of small partitions. Figure 3 graphs the number of actants (actors and artifacts) and of actors alone for the largest 86 partitions: the tail is much longer for the full set of 234. The distribution of actors roughly follows that of the total actants. As discussed later in this paper and in [24], the largest partition is centered around the very-high degree Tapped In Reception chat room (where most participants enter when they log in), the help desk volunteers who greet newcomers and direct them to their destinations, and the closely associated

ID	% Actant	Actors	Discs	Files	Chats	Avg Wdeg	Avg Path	Par Wdeg	CommType	Activity Type
0	32.1%	7659	1787	838	968	359.6	3.07	816387	Help Desk & Unsponsored	orientation, ASOs
4	5.9%	660	1016	250	131	219.9	3.90	37223	Tenant	groups, PD
15	4.7%	535	869	176	66	437.2	3.82	49945	Unsponsored	groups, PD
66	3.8%	169	1011	105	49	139.5	4.78	10519	Unsponsored	classes
76	3.2%	584	452	59	18	74.5	4.82	25392	Unsponsored	classes
14	3.0%	604	187	190	64	227.6	5.94	241533	Tenant	classes, experiment
52	2.7%	330	431	166	26	187.5	8.10	16254	Unsponsored	groups, PD
20	2.6%	331	273	207	89	67.9	4.55	16050	Tenant	groups, PD
33	2.4%	390	189	250	23	24.9	5.54	6270	Tenant	groups, PD
180	2.3%	443	268	85	25	44.5	6.32	16061	Unsponsored	groups
1	1.7%	59	275	240	33	229.0	3.51	1371	Tenant	classes
22	1.7%	116	85	356	20	105.9	3.73	2948	Unsponsored	groups
56	1.5%	183	227	82	35	311.3	4.20	4262	Tenant	groups, PD
12	1.3%	146	166	119	39	172.5	5.04	23133	Unsponsored	classes
75	1.3%	148	195	92	23	105.3	5.08	17019	Tenant	classes, PD
13	1.3%	236	44	61	109	6387.5	3.91	323835	Tenant	groups, classes
54	1.0%	174	135	25	29	282.7	4.16	52110	Unsponsored	groups, PD
38	1.0%	153	51	77	80	493.6	3.64	108829	Unsponsored	classes
176	1.0%	236	43	50	15	54.6	2.95	3243	Unsponsored	classes

ID = modularity class ID from Gephi; % Actant = percentage of total actants in the analysis (broken down in next four columns); Avg Wdeg = average weighted degree of actants in the partition (a measure of activity normalized by actants); Avg Path = average path length between any two actants in the partition (a measure of network cohesiveness); Par Wdeg = weighted degree of the edges between the partition and all other partitions (a measure of level of bridging activity); ASO = After School Online; PD = Professional Development.

Table 1. Top 19 partitions found, with selected compositional statistics and sociometrics.

set of public rooms within which public After School Online (ASO) events took place. We omit this partition from a few of our graphs in order to visualize finer structures in the other partitions. At the other extreme, 86 partitions consist of only 0.01% of the artifacts, many with one actor and one or two artifacts at the tail. These are individuals who may have interacted with others, but are more closely associated with artifacts in their offices than with partitions that involve other actors. These small partitions also produce outliers when a single individual is involved with a relatively large number of artifacts.

4.3. Distribution of parameters across sizes

We wanted to see whether parameters of interest varied across the size of partitions—that is, do the characteristics of embedded cohesive subgroups change as they get smaller? For these analyses we used the sample of every 10^{th} partition.

Some of the parameters we examined did not vary in a systematic way as a function of size. Typically there is a lot of variation, with outliers punctuated throughout the size distribution. For example, Figure 4 shows the average weighted degree, with partitions sorted by number of actors (shown in the x axis). This provides a visualization of the normalized amount of activity per actor, as each event in the data becomes one increment to the weight on an edge in the graph. While there is wide variability, no clear size-related trend can be seen. This result suggests that the extent to which people (and artifacts) participate does not vary as the size of the cohesive subgroup. Outliers shown as spikes in the graph will need to be explained by some other characteristics of the subgroup.

We also examined weighted degree of partitions (Par WDeg in Table 1). This sociometric treats entire partitions as nodes and computes the weighted degree of edges crossing partitions. It provides a measure of how much bridging or extra-partition activity there is. We expected that smaller partitions would be more "permeable" in that participants are more likely to go outside of their small partition to engage in other activities. Figure 5 suggests that this is not the case: the permeability of groups corresponding to spikes in Figure 5 will have to be explained by other factors.

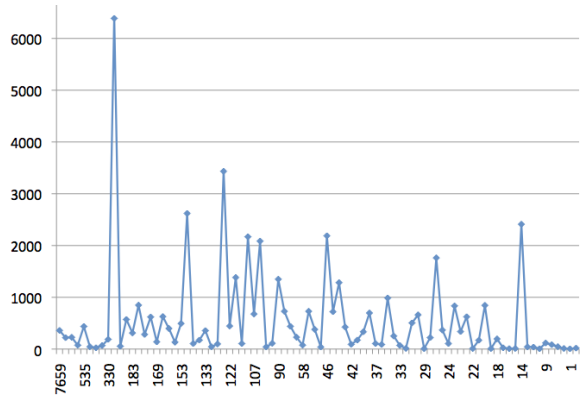

Figure 4. Average weighted degree of actants sorted by actor size of partition

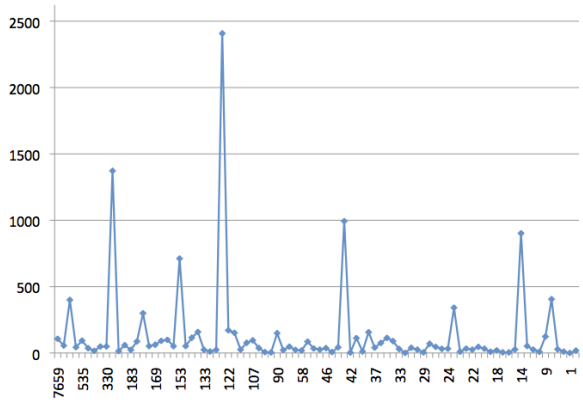

Figure 5. Weighted between-partition degree normalized and sorted by number of actors

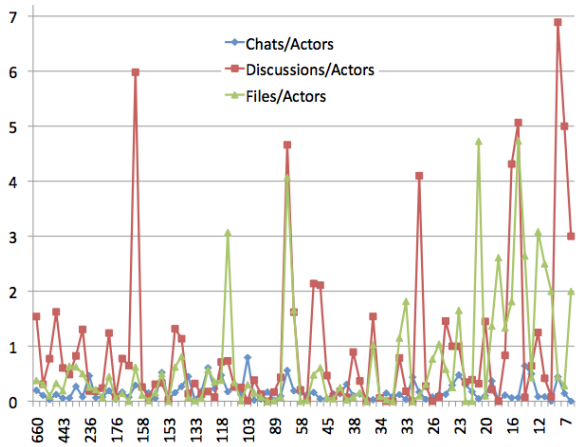

Figure 6. Artifact to Actor ratios sorted by actor size of partition

(Sensible interpretations of outliers have been found, but page limits preclude reporting those results here.)

However, other metrics show trends. Figure 6 shows the number of chats, discussions, and files in each partition, normalized by the number of actors in the partition. To make details visible, this graph excludes the largest and three smallest partitions in the sample. The three smallest partitions were removed from the sample because they had very large figures for discussions that obscured the other patterns. The graph tells us whether the per-actor availability of each of these three artifact types varies across these groups as they change in size. There appears to be no trend for chats: they are used equally across all sizes. Discussions increase for very small groups (consistent with our observation that in some small groups a few people are sharing several discussions, while many dozens may share discussions in larger groups), but there is also much variability in the discussion/actor ratios for larger groups. A trend towards greater file associations per person in small groups is also discernable. We have noticed that some small groups operate asynchronously (files and discussions), perhaps because they do not have either the critical mass or the formal mandate for synchronous sessions.

Finally we found some trends that are expected for graph-theoretic reasons. As partition size decreased, the density of the graph went up and the average path length went down. These results are expected because the number of potential links increases as the square of the graph size, while the capacity of an individual to form associations with large numbers of people does not change.

4.4. Tenant versus Unsponsored

The analysis discussed in the next section classifies partitions as being driven by paid Tenant organizations versus Unsponsored. Do Tenant versus Unsponsored activities differ from each other? To begin to answer this question, we compared the 8 Tenant versus the 10 Unsponsored partitions in the top 2-19 partitions, that is, partitions of size 1% or larger excluding Partition 0, as this large reception/ASO partition has its own unique characteristics. (The long tail was excluded to avoid confounding with size differences, with one exception below.) The number of human actors considered are similar: 2611 actors in the 8 Tenant partitions versus 2886 in the 10 Unsponsored ones (see also Table 1), so differences are not merely due to different number of actors.

The use of media is similar in the large partitions. Respectively for Tenant and Unsponsored: 50.5% vs 59.6% of artifacts are discussions, 34.7% vs 29.6% are files, and 14.9% vs 10.8% are chats. There is a slight trend towards synchronous chats in the tenant groups and asynchronous discussions in the unsponsored groups. When we look instead at the sample of every 10th partition to include the long tail of small partitions (again excluding partition 0), the balance shifts

towards asynchronous media in the Unsponsored partitions. Respectively for Tenant and Unsponsored, usage is 17.0% vs 45.8% discussions, 11.6% vs 28.6% files, and 71.4% vs 25.6% chats. These results are consistent with the trends seen in Figure 6 discussed in the previous section, and the fact that tenant organizations can provide the coordination needed to schedule synchronous chats.

Sociometrics show that Tenant and Unsponsored groups are similar, but suggest higher levels of activity in Tenant partitions. The average weighted degree of Tenant partitions is 946.65, much greater than 199.24 in the Unsponsored ones, but since the average degrees are very similar, 4.76 vs 4.62, this can be attributed to the greater use of chats in Tenant partitions, as chats generate more events. Finally, the average path lengths in the two partitions are very similar: 4.58 Tenant vs 4.74 Unsponsored. The primary results in both size-related and tenant/unsponsored analyses are a shift towards asynchronous media as groups get smaller, coupled with a surprising lack of other differences across group sizes.

5. Interpretation of Partitions

We interpreted selected partitions to see whether they make sense as "communities" or other kinds of collective activity within the TI network. The interpretation was done by sorting actants within each partition by measures of importance (degree, weighted degree, eigenvector centrality) and determining the affiliations of the highly ranked actants; by examining network structures to detect internal structure and use of media; and by examining the contents of chats and discussions in the most active settings. We also relied extensively on the SRI co-authors' knowledge and records (e.g., email archives) concerning activities during the study period.

This analysis provides a form of validation of the utility of modularity partitioning of an associogram. One approach to validating "community detection" algorithms is to generate artificial networks within which pre-identified communities have been embedded using a parameter to adjust the amount of overlap [12], and see whether the algorithm can find them. Such validation has been done with the Blondel et al. algorithm [3], but here we address a different question. How do we know that the partitions created by the modularity-partitioning algorithm can be taken seriously as potential communities or groups? We do not know or want to assume in advance what communities exist within Tapped In, because this is part of the question we are asking. However, some of us have experience running the TI network and know who the tenants and active persons were. So our approach to validation is an interpretative one: we examine the partitions generated by the algorithm, and see whether we can make sense of each partition in terms of what we know about activities within the TI network. While most of the resulting partitions were easy to interpret and not surprising, many results were not predicted. However, most of the unpredicted results were also validated in terms of what is known about the community and through examining user content. The next section interprets the largest partitions, those comprised of at least 1% of the actants, and the sample of every 10th partition of the remainder.

5.1. Partitions of 1% or more of the actants

Table 1 summarizes the features of the 19 largest partitions. Interpretation of most of these partitions was straightforward. The largest partition, with 32% of the nodes (11252 nodes), encompasses the Tapped In Reception and After School Online (ASO) community. The TI reception room has by far the highest unweighted (18,809) and weighted degrees (2,511,037) and betweeness centrality (0.675) in the context of the whole network. All of the other high ranked rooms were dedicated to ASO events. We interpret this partition to be the "entry portal" of Tapped In, where volunteer HelpDesk staff greet "newbies", and where members who primarily engaged only in ASO events interacted [9].

Eight of the top partitions were centered on the activities of 7 unique tenants (one tenant was predominant in two partitions). These partitions ranged from 450-2057 nodes and 3185-14439 edges. During the time frame examined, there were 11 tenants. As the SRI team was working closely with tenants and knew they were actively encouraging participation and preparing supporting artifacts, we expected that tenants would account for a high percentage of the larger partitions. For each partition associated with a tenant, we found people, discussions, or certain group rooms that we expected to be in the partition, making it identifiable. However, sometimes a detected tenant partition did not contain important actants (such as a key actor) known to be associated with the tenant. This spurred us to look for another partition containing this actant. In one case, we identified at least 2 other smaller partitions associated with a large-partition tenant. In another case, we noted the main structures of a tenant organization (3 sites and the headquarters) within a detected partition and thought it was complete, but later discovered an additional, smaller partition also associated with the tenant. Thus, we found that tenant activity was not always aggregated in one monolithic partition, but rather distributed across smaller activity structures that were not visible to us until now.

While it is not surprising that unplanned spin-offs or facets of planned tenant activity happen, we were not always aware of it, and there is value in discovering this empirically. The discovered structures reflect the reality that not all members of an organization interact frequently, and small clusters of people often work closely together and somewhat separately from other clusters. Indeed, it can be inefficient for all members to be connected [17]. Future analysis could examine the adequacy of information flow between the separate cohesive subgroups of an organization, or across organizational boundaries [5].

Of the remaining (non-tenant) top partitions, 5 were online classes and 5 were online groups supporting teachers in professional development activities. The 5 online classes were all were all university-level classes led by Tapped In members who we categorize as "lone rangers": Tapped In members who independently and successfully established an online place in Tapped In where they could work with students or with like-minded colleagues from their own institution or elsewhere. Of the online groups, 4 were led by a lone ranger from an organization trying to support teachers. The remaining group was led by a successful online leader and long-time Tapped In member who was consulting for two organizations, one of which was a Tapped In tenant. We see those two organizations and their work together in one partition because of this very active leader. While we expected to see tenants in the top partitions, we were not aware of how active these other non-tenant classes and groups were, and only learned of their extensive activity through the present analysis.

5.2. Partitions of less than 1% of actants

Next, we consider a sampling of 21 partitions from the 215 smaller partitions, those that each contain less than 1% of the nodes. These 21 small partitions represented a "long tail" in that they varied widely in their activities and purpose. The smallest 6 partitions were primarily individuals experimenting with artifacts in the system (e.g., testing out the discussion board or downloading a file), and contained fewer than 5 nodes, including both actors and media. Unlike sociograms consisting only of actors, associograms cluster people with artifacts, enabling community leaders to discover what people are doing even if they are not interacting with other people.

The remaining 15 small partitions could be understood as different kinds of "groups". These partitions included the following: 3 affiliated with tenants, 3 online classes (2 at the university level and one in equine science), 4 informal discussions (chat or threaded discussion), 1 file-sharing group for people who worked together face to face, 1 group of 3 people who downloaded the same documents but did not have any other interaction (so their link was strictly through the artifact), and 3 associated with K-12 student groups. (While data on K-12 students was not included, we can see adult interactions with artifacts/media clearly associated with K-12 settings in K-12 student group rooms.) Some of these partitions included actors associated with a geographical region (e.g., state or city), perhaps supplementing face-to-face meetings. Others included actors widely distributed across the U.S. or the world, likely using the space as their primary means of interaction. Not surprisingly, the partitions affiliated with tenants were relatively larger partitions, with 40, 68, and 105 nodes respectively. Prior to this analysis, we suspected that many people were using Tapped In to support activities with peers and students, and indeed, this sample revealed a wide variety of diverse activities.

6. Discussion and Conclusions

The analysis reported here improves on the analysis reported in [22] in that (a) the present analysis uses a new version of Gephi (0.8.1) with a corrected implementation of [3] to use weights for determining the partitions; (b) the analyses and results in section 4 on partition distribution are entirely new; (c) the process by which partitions were interpreted in section 5 differs, involving community facilitators; and (d) there was further data cleaning to deal with small problems found. Although the number of partitions found differed (due to (a) and (d)), it is comforting that the major partitions found in the two analyses are quite similar in their interpretations.

The analysis explored what we could learn from structures detected automatically by modularity partitioning of associograms. We confirmed that human facilitators could interpret detected structures as meaningful. We expected (and confirmed) that tenants were engaged in high levels of activity, but learned that this activity was sometimes distributed into clusters. We knew that many other groups were using Tapped In informally, but did not have a good handle on these "unsponsored" groups; this analysis helped us understand the variety and frequency of various types of non-tenant-affiliated online activities.

6.1. Implications for researchers

As researchers, we need tools to help us understand what is occurring in online networks and to then interpret what the results mean [19]. For many years, without tools, our research was limited to studying the groups we knew—in this case, mostly the tenant organizations. We were aware that we did not

know whether the groups we were studying were representative of the rest of the community, but we did not have a way to address this problem. For the first time, we see the structure of this heterogeneous network. We found a large network consisting of the entry portal and public events, the "transcendent community" [10] within which other activity and groups are embedded. Sponsored activity was found as expected, but not always in one setting. A large range of unsponsored activity was also found, some comparable in size to the sponsored groups but also in a diverse long tail of smaller groups and individuals. Although there is more of everything in larger groups, it was surprising to find that per-person measures of activity or bridging did not vary as a function of size, but rather spiked in an idiosyncratic manner. The only trend was a shift towards asynchronous media in smaller and unsponsored groups.

The methods discussed in this paper pose the data in a manner that leads us to ask questions we might not otherwise have thought of. Much more analysis is needed to fully understand what was occurring. For example, can sociometrics of connectedness between partitions potentially be employed to help detect partitions that are related, specifically to find more "parts" of a tenant organization (e.g., if a tenant has multiple, connected clusters)? Identifying connected subgroups within an organization may further our understanding of when and how organizations break into smaller groups in online activities.

Further automated sociometric and content analysis tools will make this work manageable. We need to develop ways to more quickly understand what certain graph structures might indicate. Improvements in cohesive subgroup detection methods are also needed. Most methods, such as the one we used, force a partition: each actor or actant can be associated with only one "community" However, in most natural systems, including the TI network, actants play roles in multiple settings. New algorithms for *overlapping* "community detection" are available, but some (e.g., [16]) do not work on bipartite graphs. We are evaluating the suitability of edge community [1] and flow compression [8] approaches for our next analysis of this data.

6.2. Implications for practitioners

Many organizations create online mentoring or professional development programs, or online adjuncts to face-to-face programs, and find they have few metrics to interpret or assess what is happening beyond basic counting or laborious analysis of the discourse that participants produce. Looking at frequency data (e.g., number of posts over time) or analyzing discussion postings at the end of a project occurs too late to do anything to make a difference. Evaluators and facilitators can only try again with the next cohort. If we can use automatically collected data to see what "invisible" work is occurring within and between online groups, we could begin to reliably identify patterns that lead to failure or success, both for individuals and entire networks. As automated analysis tools continue to be developed and refined, how might they help group leaders support, learn, and collaborate with one another more effectively in online communities [19]? In conducting the current analysis, we identified ways that the results of automated analyses might have helped us in the practice of supporting an online community. For example, we could have reached out to low-degree members ("newbies") appearing in the Reception and ASO partition, asking whether they needed assistance (perhaps with automatically-generated emails, since this is a large partition). For people who seemed to be leaders (e.g., actants with high degree or centrality metrics), we could have reached out with individual emails to ask how we could better support their efforts. For groups that appeared to use a dominant form of communication (e.g., threaded discussion), we could have reached out to ask whether they knew about other tools (e.g., chat, file sharing) available in the environment, or wanted to learn more about how they might these tools. Finally, knowing the pivotal role in of bridges in the diffusion of information [4], we could have targeted dissemination of key information to individuals or groups where we saw evidence of bridging or extra-partition activity.

We need to have tools that tell us how much members of online networks interact, with whom, concerning what, and eventually link the interactions to outcomes. We need this kind of information, not just in small slices of time for small samples of people, but aggregated across whole networks over extended periods of time. We need to look at a variety of groups, not just "interesting" groups, so that we can know what is normal and dysfunctional, as well as exceptional: how much online interaction it takes to make a difference; what the trajectory is for groups as they begin to work online; which people are isolated and why; and how to help new people or those isolated become more involved. This work is the beginning of this vision.

8. Acknowledgments

This work was supported by NSF Award 0943147. The views expressed herein do not necessarily represent the views of NSF. The authors thank Nathan Dwyer and Devan Rosen for their prior collaborations.

9. References

[1] Y.-Y. Ahn, J. P. Bagrow and S. Lehmann, Link communities reveal multiscale complexity in networks, *Nature*, 466 (2010), pp. 761-765.

[2] M. Bastian, S. Heymann and M. Jacomy, Gephi: An open source software for exploring and manipulating networks, *International AAAI Conference on Weblogs and Social Media*, 2009.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, http://dx.doi.org/10.1088/1742-5468/2008/10/P10008 (2008).

[4] R. Burt, The network structure of social capital, in B. M. Staw and R. I. Sutton, eds., *Research in organization behavior, Volume 22*, JAI Press, Greenwich, CT, 2000.

[5] V. Buskens and A. Van de Rijt, Dynamics of networks if everyone strives for structural holes, *American Journal of Sociology*, 114 (2008), pp. 371-407.

[6] M. Castells, *The Internet Galaxy: Reflections on the Internet, Business, and Society*, Oxford University Press, 2001.

[7] A. P. Cohen, *The Symbolic Construction of Community*, Routledge, New York, 1985.

[8] A. V. Esquivel and M. Rosvall, Compression of flow can reveal overlapping-module organization in networks, *arXiv*:1105.0812v4 [physics.soc-ph] (2011), pp. 1-10.

[9] U. Farooq, P. Schank, A. Harris, J. Fusco and M. Schlager, Sustaining a community computing infrastructure for online teacher professional development: A Case Study of Designing Tapped In, *Computer Supported Cooperative Work*, 16 (2007), pp. 397-429.

[10] S. Joseph, V. Lid and D. D. Suthers, Transcendent Communities, in C. Chinn, G. Erkens and S. Puntambekar, eds., *The Computer Supported Collaborative Learning (CSCL) Conference* 2007, International Society of the Learning Sciences, New Brunswick, 2007, pp. 317-319.

[11] H. Kim, G. J. Kim, H. W. Park and R. E. Rice, Configurations of relationships in different media: FtF, email, instant messenger, mobile phone, and SMS, *Journal of Computer Mediated Communication*, 12 (2007), pp. 1183-1207.

[12] A. Lancichinetti, S. Fortunato and F. Radicchi, Benchmark graphs for testing community detection algorithms, *Physical Review E*, 78 (2008), pp. 046110-1-5.

[13] B. Latour, *Reassembing the Social: An Introduction to Actor-Network-Theory*, Oxford University Press, New York, 2005.

[14] C. Licoppe and Z. Smoreda, Are social networks technologically embedded? How networks are changing today with changes in communication technology, *Social Networks*, 27 (2005), pp. 317-335.

[15] S. Martin, W. M. Brown, R. Klavans and K. Boyack, OpenOrd: An Open-Source Toolbox for Large Graph Layout, *SPIE Conference on Visualization and Data Analysis* (VDA), 2011.

[16] G. Palla, I. Derényi, I. Farkas and T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, *Nature*, 435 (2005), pp. 814-818.

[17] W. R. Penuel and M. Riel, The new science of networks and the challenge of school change, *Phi Delta Kappan*, 88 (2007), pp. 611-615.

[18] K. A. Renninger and W. Shumar, *Building Virtual Communities: Learning and Change in Cyberspace*, Cambridge University Press, Cambridge, 2002.

[19] M. Schlager, U. Farooq, J. Fusco, P. Schank and N. Dwyer, Analyzing online social networking in professional learning communities: Cyber networks require cyber-research tools, *Journal of Teacher Education*, 60 (2009), pp. 86-100.

[20] M. Schlager, J. Fusco and P. Schank, Evolution of an Online Education Community of Practice, in K. Renninger and W. Shumar, eds., Cambridge University Press, *Building Virtual Communities*, 2002, pp. 129-158.

[21] L. S. Sproull and S. B. Kiesler, *Connections: New ways of working in the networked organization*, MIT Press, Cambridge, MA, 1991.

[22] D. D. Suthers and K.-H. Chu, Multi-mediated community structure in a socio-technical network, *Proceedings of LAK12: 2nd International Conference on Learning Analytics & Knowledge*, April 29 - May 2, 2012, Vancouver, BC., ACM, New York, 2012.

[23] D. D. Suthers, N. Dwyer, R. Medina and R. Vatrapu, A framework for conceptualizing, representing, and analyzing distributed interaction, *International Journal of Computer Supported Collaborative Learning*, 5 (2010), pp. 5-42.

[24] D. D. Suthers and D. Rosen, A unified framework for multi-level analysis of distributed learning, *Proceedings of the First International Conference on Learning Analytics & Knowledge*, Banff, Alberta, February 27-March 1, 2011, 2011.

[25] D. Tapscott and A. D. Williams, *Wikinomics: How mass collaboration changes everything*, Portfolio, New York, NY, 2006.

[26] F. Tönnies, *Community and Civil Society* (J. Harris & M. Hollis, Trans. from Gemeinschaft und Gesellschaft, 1887) Cambridge University Press, Cambridge, United Kingdom, 2001.

[27] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, New York, 1994.

[28] B. Wellman and M. Gulia, Net surfers don't ride alone: Virtual communities as communities, in M. A. Smith and P. Kollock, eds., *Communities in cyberspace: Perspectives on new forms of social organization*, Routledge, London, 1999, pp. 167-194.

[29] B. Wellman, A. Quan-Haase, J. Boase, W. Chen, K. Hampton, I. Diaz and K. Miyata, The social affordances of the internet for networked individualism, *Journal of Computer-Mediated Communication*, 8 (2003).