Hypergraph Characterization of Small Groups

Ankit Sharma[a], **Himanshu Kharakwal**[b], **Abhishek Chandra**[a], **and Jaideep Srivastava**[a]

[a] Department of Computer Science & Engineering, University of Minnesota, USA 55455; [b] LNM-IIT, Rajasthan, India

This manuscript was compiled on July 6, 2018

Understanding groups is of prime importance in sociology as well as psychology. In this research we focus on small groups. We aim to empirically answer the questions of segregating small groups from plethora of social groups (macro level analysis) as well as distinguishing characteristics among small groups (micro level analysis). We use temporal hypergraphs as the model to capture the self assembling small groups over time. Various cross-sectional as well as longitudinal hypergraph metrics are developed to characterize as well as interpret dynamics of small groups. We observe that small groups distinctly follow a Log-normal cardinality distribution unlike other kinds of social groups like interest based groups. Group variability metrics are developed for studying the inter-groups interactions. Cross-sectionally, these metrics result in a good Weibull type fit. Over time the group variability of research groups increase and the members of the group increasingly start participating in external groups, but patent groups are more covert with a low group variability in general. Task oriented online gaming groups tend to self assemble into good chemistry groups less distracted by external participation. To further capture the average individual attention, a group level attention metric is developed. This metrics follows a log-normal cross-sectionally across the datasets. However, temporal attention metric seems to be decreasing over years in research teams but increases in task oriented and highly competitive gaming teams and for that matter is quiet high in (covert type) patent teams, across years. Interestingly, these metrics tend to successfully track fluctuations in Enron dataset during the time of scandal. Overall in this research we leverage hypergraph, which naturally model self-assembling small groups, to develop interesting novel metrics, to study small groups at totally different level of granularity which naturally takes into account the network effects.

hypergraph | small groups | network science | social science | social networks

Social groups have been studied extensively across social sciences (?) and have found extensive real life applications (?). Internet generated online space has allowed for new types of group interactions more than ever before. These new interactions provide minute by minute traces of group interaction data (?) making their study possible in a manner which is much more fine grained than ever before. Groups have been defined in several ways (?) in social sciences. More often the primary dimension to categorize groups is according to their sizes. *Small groups* can typically contain 3 to 20 members (?) whereas medium sized *guilds* have less than 200 members (?). Finally, their are larger arbitrary size communities which can have hundreds or even thousands of members as part of them (?). Social groups co-exist within the social network between the social actors. This network of actors influences the group dynamics which in-turn affects the network evolution. Large part of literature within network science is dedicated to finding groups and its evolution (?). However, their research lacks any proper definition of groups and resulting in arbitrary sizes ranging all the way from small groups to huge communities (?). On contrary our research focuses on well defined small groups and we propose the notion of network of groups. Small groups occur naturally or are formed as teams to jointly carry out tasks. We believe that small group phenomena is highly distinct from that of larger guilds or communities and therefore, are governed by different set of theories. The group structure is already given in form of the group interaction data. We focus more on building models that capture the network of groups notion and studying the group dynamics over this network. We propose the hypergraph model as network of groups and build several metrics to infer interesting characteristic group behavior.

1. Model

Consider the scenario where we have a set of n individuals or social *actors*: $V = \{v_1, v_2, ..., v_n\}$. These *actors* self assemble themselves into *groups* to perform tasks at hand or gather for an event. A *group* therefore, is a subset of all the *actors*. We have a collection of m such groups observed in past, denoted by $G = \{g_1, g_2, ..., g_m\}$ where $g_i \subseteq V$ represents the i^{th} *group*. Cardinality $c_i = |g_i|$ of a group is the number of actors part of it. Membership of a *group* can change over time. An *actor* can leave or join a *group*, resulting in changes in *group* membership. When two *actors* work or gather together in the same *group* they develop social tie. These social ties therefore, become the edges in the social *network of actors* (NOA). NOA is a graph, $N_a = (V, E)$ where $E = \{e_1, .., e_w\}$ are the dyadic edges defined over vertex set V. Moreover, the *actors* that are the part of multiple groups act as ties between groups resulting in a *network of groups* (NOG). NOG is a hypergraph (?) represented as a set $N_g = (V, G)$ with G as the hyperedges over the vertex set V. Given a past history of *groups* we therefore, can construct both the NOA and NOG graph theoretic models for this data. Most of the past research has looked at NOA model, however, in this research we shall stress on the NOG model and build adequate metrics to capture group behavior of interest.

Significance Statement

Authors must submit a 120-word maximum statement about the significance of their research paper written at a level understandable to an undergraduate educated scientist outside their field of speciality. The primary goal of the Significance Statement is to explain the relevance of the work in broad context to a broad readership. The Significance Statement appears in the paper itself and is required for all research papers.

Please provide details of author contributions here.

Please declare any conflict of interest here.

[1] A.O.(Author One) and A.T. (Author Two) contributed equally to this work (remove if not applicable).

[2] To whom correspondence should be addressed. E-mail: author.twoemail.com

Notice that $N_g = (V, G)$ is the complete set of groups (hyperedges) and actors (vertices), observed over a period of time $[0, T]$. However, at each time instance $t \in [0, T]$, only a subset of the groups, $G^t \subset G$, (and the corresponding actors $V^t \subset V$), might only be active. We refer to G^t as the *snapshot* (of NOG or hypergraph).

We also categorize groups into *core* groups which do not contain any subgroups within them not part of any other groups and all other groups which might contain subgroups within them are the *expanded* group. Furthermore, a *core* group along with all the *expanded* groups it is part of, constitute a *family*.

2. Results and Discussion

Our research is conducted using ten different datasets. These data sets are all social group oriented however, they vary in the size, type as well as intensity of social interactions. Refer *Materials and Methods* for dataset details. In the following sections we will describe results for four different kinds of metrics built using the hypergraph model described in the previous section. In next section we initiate with a discussion on the macro-level question of: How to distinguish small group data from general social groups? In the remaining sections we focus only on the small group datasets. The stress would be on comparing different small group data sets, as to what characteristics are variable or invariable. Therefore, resulting in various interesting conclusion on behavior of within small groups.

A. Distinguishing features of Small Groups. The ten datasets that we considered in our research fall in two major genre. In one genre we have groups that are set of individuals sharing common interests and the second type consists of groups which are set of individuals that interact together as a team to perform various tasks. Our aim is to develop metrics that can very well distinguish the different genre empirically. For this we consider, the first metric of *Group Size* $(\mathbf{\Psi_0})$: the number of actors or members part of a group. Figure 1 show the Probability Mass Functions for Group Size. We observe that all these datasets have a heavy tail (cloud of point at tail). Visually, we can further observe that there are two distinct categories of datasets. Interest group datasets, which include Friendster (FS), LiveJournal (LJ) and Youtube (YT), have body which is pretty much a straight line. Other datasets, which are more team oriented, have a body that is not a straight line. In fact we try to fit log-normal distribution to all the ten datasets, and observe that log-normal in general fits the team oriented small group datasets.

Heavy tail cloud in PDF causes difficulty in fitting, therefore, we use the Complimentary Cumulative Distribution Function (C-CDF) curve for fitting our hypothesized distributions. Figure 2 show the C-CDF for group sizes. We can observe in Figure 2 that the body is very well fitted by the *log-normal* distribution whereas we find (by visual analysis) that a large portion of the tail is well fitted by *power-law* (i.e. a line shown as dotted black line). Typically this transition from the *log-normal* to *power-law* occurs at different points for various datasets. Power-law fit suggests the prominence of preferential attachment in the larger size groups in tail. Therefore, a large group attracts more individuals to become part of it in proportion to its current size. Rich get richer phenomena is at play in large size groups.

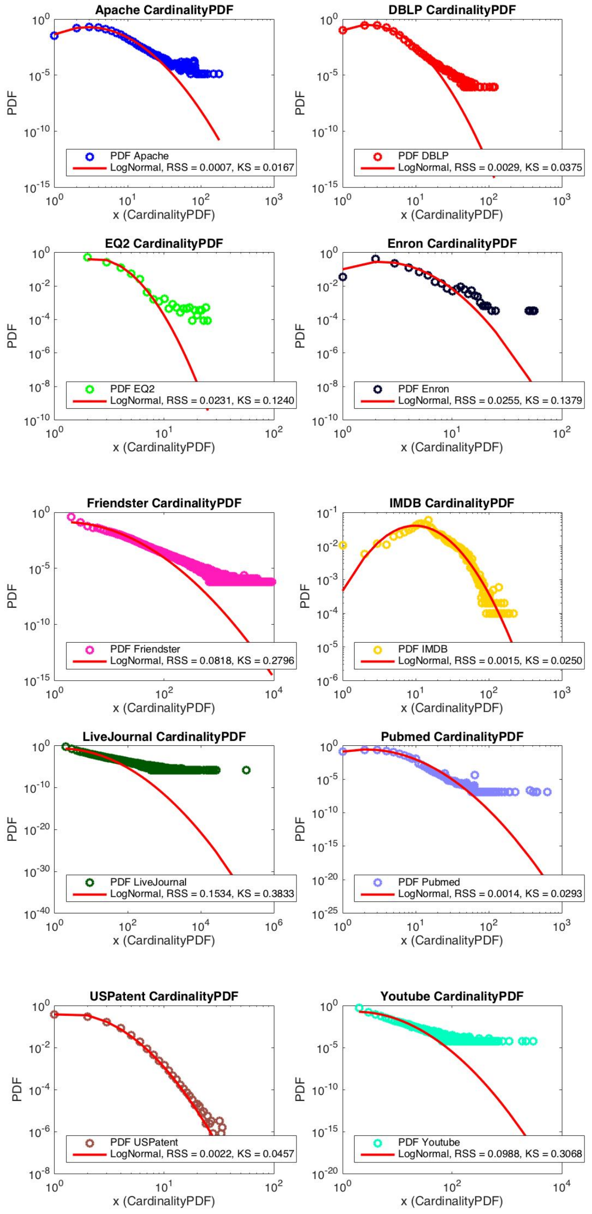

Fig. 1. Log-Log plot for Probaility Mass Function (PDF) for the group size or cardinality. X-axis is the group size and Y-axis is the P(X=x) i.e. probability of finding size equal to x. Body of the curve is well fitted by a *log-normal* ditribution.

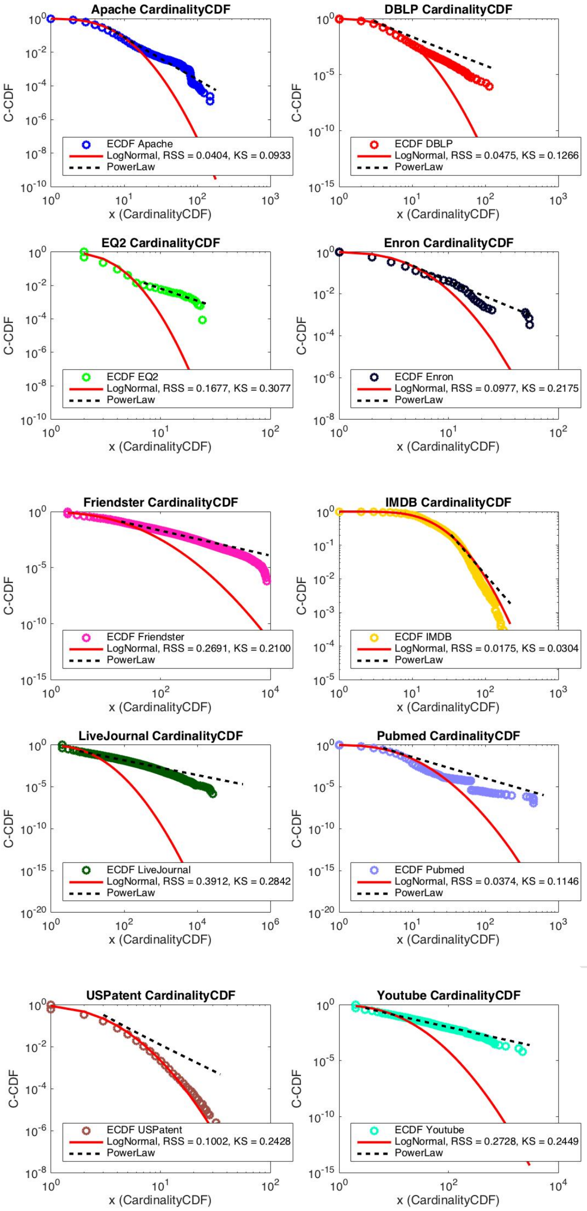

Fig. 2. Log-Log plot for Complementary Cumulative Distribution Function (C-CDF) or Survival Function for the group size or cardinality. X-axis is the group size and Y-axis is the P(X>x) i.e. probability of finding size greater than x. Body of the curve is well fitted by a *log-normal* ditribution while a major portion of the tail is *power-law*.

B. Characteristic Group Sizes of Small Groups. Among the various kinds of small group datasets, *group size* can further help us distinguish their nature. We observe that the peak of the lognormal fit in Figure 1 occurs at different cardinalities. For example for movies in IMDB data a team of ten individuals seems quiet prominent where as in research teams in DBLP or Pubmed sizes between two to four seems most popular statistically. One can relate this most prominent size (peak in the log-normal fit), as a reflection of the kind of group interaction involved. Movie making is complex task involving various individuals of varying skills therefore, ten people seems to be some kind of an ideal size for a successful movie making. On the contrary, research publication tasks requires intensive interaction among like minded or skilled individuals. Possibly, intense interaction is more easily done in smaller teams, therefore, making small size teams more characteristic in research tasks.

C. Characterizing Inter-Group Interactions. In the previous section we characterized an individual group specific attribute, namely the group size. However, the real-world groups are in constant interaction via the common members. In this section we shall focus on understanding the interaction of a group (hyperedge) with it's neighboring groups (overlapping hyperedges within NOG). While capturing these interactions we aim to understand the repercussions of these interactions on the individual group's behavior. We investigate the inter-group interactions using three different metrics.

The first metric is that of *Group Variability* which captures the number of incident hyperedges to a given group's hyperedge within the NOG. In order to remove the bias of the larger groups, we further normalize by dividing it by the given group's size. More formally let us define the set G_g associated with the given group hyperedge g such that $G_g = \{e|e \cap g \neq \emptyset, e \subset G, e \neq g\}$ i.e. set of all the incident hyperedges. We can then define the metric of *Group Variability* ($\mathbf{\Psi_1}$) as follows:

$$\mathbf{\Psi_1}(g) = \frac{|G_g|}{|g|} \qquad [1]$$

where g is the given group hyperedge.

As can be observed from the Figure 3 that at group level this metrics is very well fitted by *weibull* distribution.

However, the group variability metric does not take into account the extent of overlap between the given group with its incident group hyperedges. To accommodate that we define another metric as follows:

$$\mathbf{\Psi_2}(g) = \frac{\sum_{e \in G_g} |e \cap g|}{|G_g||g|} \qquad [2]$$

As this metric captures the bridging capability of the given group we call it *Overlap-based Hyperedge Bridging Capital (HBC-Overlap)* ($\mathbf{\Psi_2}$). It is simply the normalized sum of fraction overlaps of the given group with other groups incident on it. Notice that $\frac{1}{|g|} \leq \mathbf{\Psi_2}(g) \leq 1$. Figure 4 show the C-CDF curves for HBC-Overlap metric. At the group level over all *log-normal* is a good fit rather than *weibull*. But both distributions fails to capture the tail.

We further more define another metric that captures the over all attention span of the group members towards the given group. For this we start by defining the degree of the vertex as the number of groups a vertex is a part of. Formally

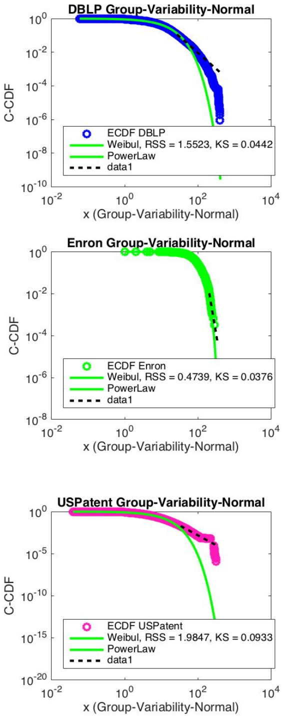

Fig. 3. Log-Log plot for Complementary Cumulative Distribution Function (C-CDF) or Survival Function for the group variability. X-axis is the group variability and Y-axis is the P(X>x) i.e. probability of finding group variability greater than x. The complete curve is very well fitted by a *weibull* ditribution.

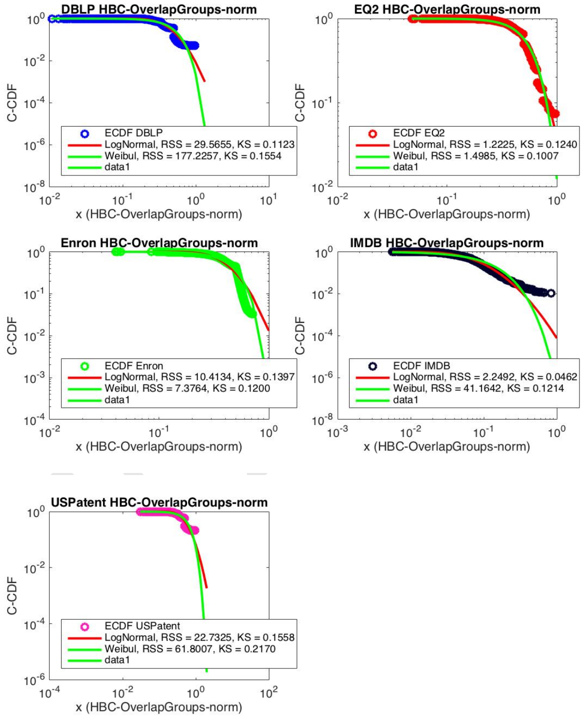

Fig. 4. Log-Log plot for Complementary Cumulative Distribution Function (C-CDF) or Survival Function for the group variablity as Overlap based Hyperedge Bridging Capital (HBC-Overlap). X-axis is the HBC-Overlap group variability and Y-axis is the P(X>x) i.e. probability of finding group variability greater than x. The whole curve is well fitted by a *weibull* ditribution. Although the body is better fitted by *log-normal* but it does not fit tail. Tail also has a strong *power-law*.

as $d(v) = |\{g|v \subset g, g \in G\}|$. Using the degree of a given actor we define another group level metric as follows:

$$\mathbf{\Psi_3}(g) = \frac{\sum_{v\in g}(1/d(v))}{|g|} \qquad [3]$$

Numerator is basically the sum of each actor's fraction attention (inverse of degree of actor vertex) span towards the given group. As this metric captures the bridging capability of the given group through its members external participation, we call it *AttentionSpan-based Hyperedge Bridging Capital (HBC-AttentionSpan)* ($\mathbf{\Psi_3}$). Notice that $0 < \mathbf{\Psi_3}(g) \leq 1$.

HBC-AttentionSpan metric shows a *log-normal* fit as a characteristic. Overall *Log-normal* fits better or very close to *weibull* when evaluated using both Root mean squared error (RSS) as well as KS statistic (KS).

Note that all metrics develped in this section so far quantify the amount of inter-group, but provides no information with respect to the diversity among the group members and among the extent of overlaps. In order to capture these diversities we develop information entropy based metrics as follows. For capturing the diversity among the actors for external collaboration, we propose *Group Variety (Actor Degree)*:

$$\mathbf{\Psi_4}(g) = \frac{-\sum_{v\in g} p(v)\ln p(v)}{\ln|g|}, p(v) = \frac{d(v)}{\sum_{v\in g} d(v)} \qquad [4]$$

where $p(v)$ is the fractional degree contribution of the vertex v towards the group variability of group g. The numerator is information entropy among the group actors ($v \in g$) with respect to their $p(v)$. Denominator is the maximum entropy achievable by a group and acts as a normalizing constant.

D. Temporal Characteristics of Small Groups. In the previous sections we have observed the groups while ignoring the temporal information associated with their data. Unlike the interest groups which are groups by association (one time activity), small groups are involved in various activities or tasks over time. Therefore, neglecting the temporal aspect, may result in only a limited cross-sectional picture which disregards the dynamics within small groups. We incorporate the temporal information by studying the metrics developed above over time (hypergraph snapshots). For this we define the temporal group metrics ($\mathbf{\Theta_i}$) corresponding to each of the (*static*) group metrics ($\mathbf{\Psi_i}$), as follows:

$$\mathbf{\Theta_i}(g,t) = \frac{\sum_{g\in G^t}\mathbf{\Psi_i}(g)}{|G^t|}, \forall i \in [0,4] \qquad [5]$$

We start with the group temporal size ($\mathbf{\Theta_0}$), whose curve is shown for all the five small group datasets in Figure 7. We observe that over years the average size of research teams have increased. This trend is shown in DBLP, Pubmed as well as in USPatent (where the task of patent is quiet similar to research tasks). On the contrary the gaming teams in EQ2 show a slightly decreasing trend in the team sizes. A possible reason could be that over time of several months the game players assemble themselves in smaller but more optimal units who perform favorable in tasks. Email communication task (in Enron) seems to be indifferent to time and there seems to be nothing like a favorable number of people to whom an email should be sent to which the organization should evolve to.

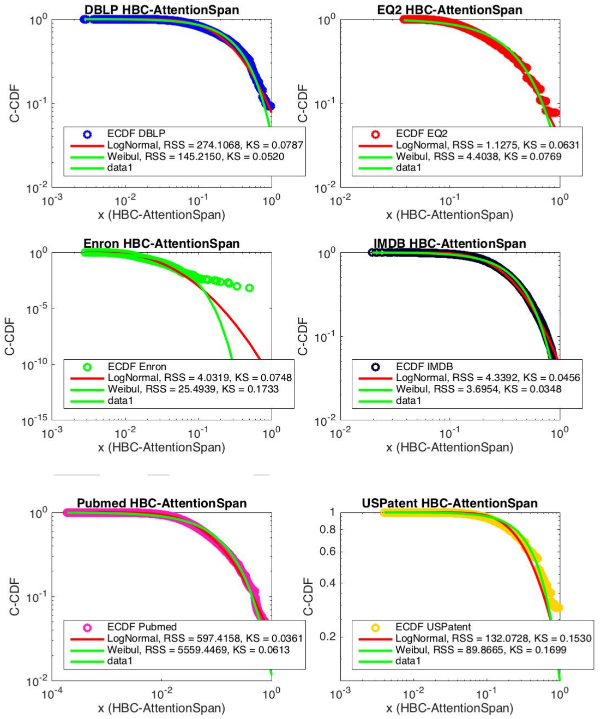

Fig. 5. Log-Log plot for Complementary Cumulative Distribution Function (C-CDF) or Survival Function for the group variablity as Attention Span based Hyperedge Bridging Capital (HBC-AttentionSpan). X-axis is the HBC-AttentionSpan group variability and Y-axis is the P(X>x) i.e. probability of finding group variability greater than x. The whole curve is well fitted by a *log-normal* ditribution.

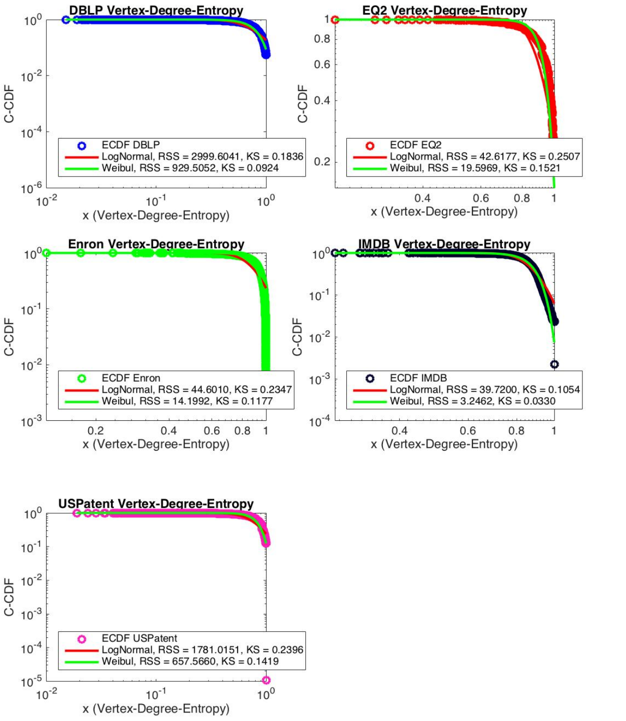

Fig. 6. Log-Log plot for Complementary Cumulative Distribution Function (C-CDF) or Survival Function for the group variety for vertex degree. X-axis is the HBC-AttentionSpan group variability and Y-axis is the P(X>x) i.e. probability of finding group variability greater than x. The whole curve is well fitted by a *log-normal* ditribution.

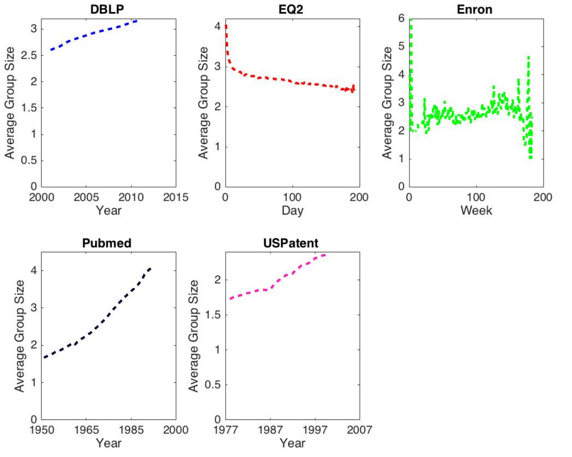

Fig. 7. Over time plot for Average Group Cardinality per snapshot. X-axis is the time snapshot t and Y-axis is the Average Group Cardinality ($\mathbf{\Theta_0}$) at time snapshot t. The whole curve shows ....

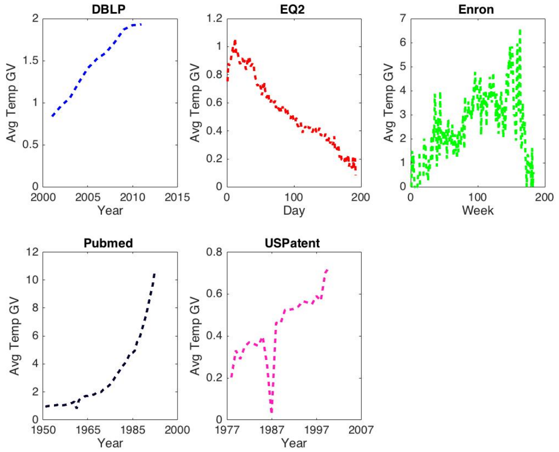

Fig. 8. Over time plot for Average Group Variability per snapshot. X-axis is the time snapshot t and Y-axis is the Average Group Cardinality ($\mathbf{\Theta_1}$) at time snapshot t. The whole curve shows ....

We shall now consider the normalized temporal group variability ($\mathbf{\Theta_1}$) its weighted version i.e. temporal HBC-OG ($\mathbf{\Theta_2}$), in tandem. As we can observe in Figure 8, for the case of research teams (DBLP and Pubmed) as well as the Enron's email communication groups , the number of other groups on an average a group's members are active in increases over the years. A very important point to notice is that these group variability value is for a particular time snapshot. Therefore, it actually reflects the multi-tasking capability of group. For example in research teams it reflects how many different external projects the group members were simultaneously active in within than snapshot (multi-tasking). It's interesting to notice that USPatent although showing an increasing trend but the values of $\mathbf{\Theta_1}$ are quiet low. A possible hypothesis that can be inferred is that task of patent making is somewhat more covert than the research publication, and groups working on patent level ideas, are more inert and deliberately avoiding external participation.

However, simply taking the count of external groups a given group overlaps with in a given snapshot does not say much about the extent of this overlap. Specially with the increase in average size of groups as well as the number of groups (see Materials and Methods), like that of research teams. For this we observe the trends in the temporal HBC-OG metric, as shown in Figure 9. We can clearly observe the dampening affect on the group variability when weighted with the overlap degrees. Enron email communication groups show lower as well as increasingly constant average external participation. This is even more observed in the case with research teams like that of DBLP and Pubmed, where the HBC-OG even shows decreasing trend over the years. This basically tells that although the average groups sizes have increased over time but large fractions of group participate externally. The other extreme could have been when external participation only happens via single common member. This basically suggests that over time high performing subgroups are formed which multi-task in various research projects possibly with various external members.

Its is interesting to notice that in USPatent data the values for HBC-OG lower to that of there small group datasets.

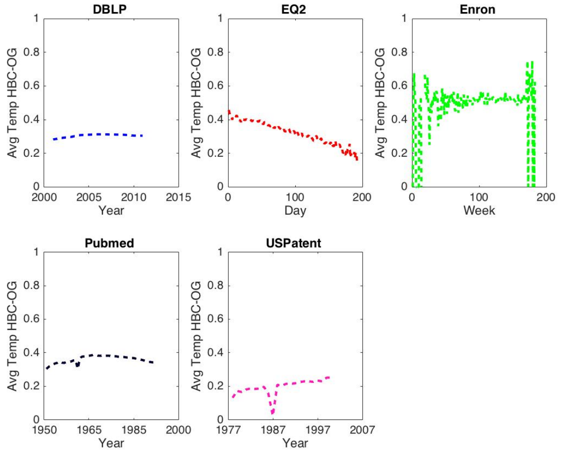

Fig. 9. Over time plot for Average HBC-OG per snapshot. X-axis is the time snapshot t and Y-axis is the Average Group Cardinality ($\boldsymbol{\Theta_2}$) at time snapshot t. The whole curve shows ....

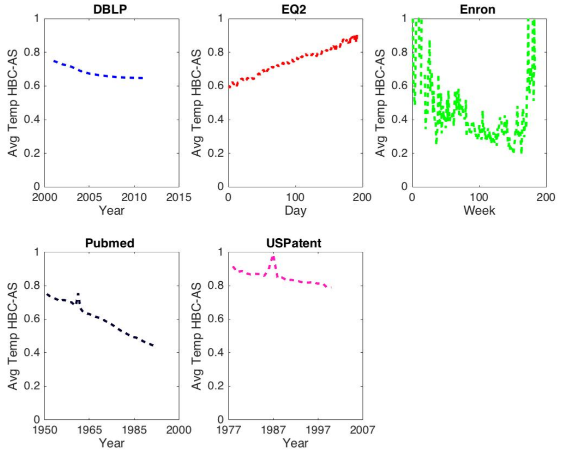

Fig. 10. Over time plot for Average HBC-AS per snapshot. X-axis is the time snapshot t and Y-axis is the Average Group Cardinality ($\boldsymbol{\Theta_3}$) at time snapshot t. The whole curve shows ....

This suggests that a large fraction of the same group works with other groups. Therefore, quiet hight overlap between groups, suggesting low encouragement to bring in new external collaborators. Which further corroborates that covert nature of patent groups.

Another thing to observe is that Enron out of all the small group datasets has the highest HBC-OG values, suggesting that overlaps between the email groups is not that high, and actors are participating even individually in various other email communications. This also reflects that email communication is also a less intense activity as compared to research or publishing patents. However, if we observe the Enron at the end weeks (when the Enron crisis happened), there are significant fluctuations in both the group variability as well as HBC-OG. Possibly relating to the change in the communications pattern between the organization's member, with the scandalous groups going covert.

EQ2 dataset has a different story again reflecting on the distinct nature of the group tasks within online games. The group variability of these groups was never too high and shows a decreasing trend. Also the extent of overlap with other groups is quiet high, suggesting the closed nature of gaming teams, which are highly selective in whom they play with. Also over time they self assemble into more inert and good chemistry groups.

HBC-AS captures the average attention of group members towards the group and is shown in the Figure 10. Over the years it seems that researchers are multitasking increasingly more and working in various groups simultaneously. Therefore, a decreasing focus in a single group activity. However, the gaming teams in EQ2 and their members, on contrary, get increasingly focused over months.

3. Materials and Methods

A. Data Preparation. In this research we have utilized nine publicly available and one proprietary data. The details of these datasets are provided in the Table 1. Only five of the datasets have temporal information available in them. Average number of active actors and groups across these five temporal

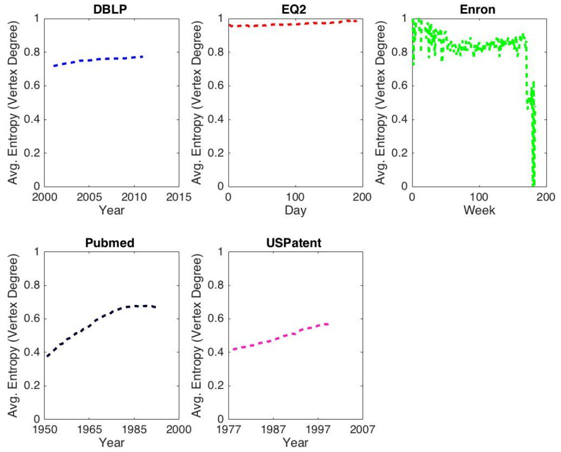

Fig. 11. Over time plot for Group Entropy (vertex degree) per snapshot. X-axis is the time snapshot t and Y-axis is the Average Group Entropy (Vertex Degree) ($\boldsymbol{\Theta_4}$) at time snapshot t. The whole curve shows ....

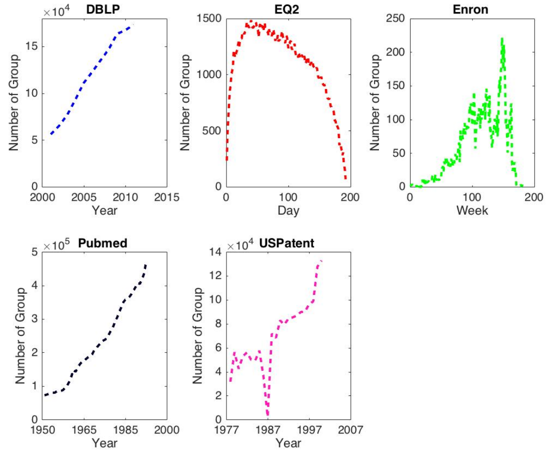

Fig. 12. Over time plot for number of active groups. X-axis is the time snapshot t and Y-axis is the Number of Active groups at time snapshot t. The whole curve shows ....

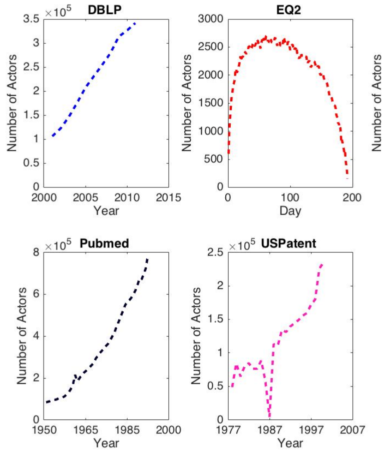

Fig. 13. Over time plot for number of active actors. X-axis is the time snapshot t and Y-axis is the Number of Active actors at time snapshot t. The whole curve shows ....

datasets are shown in Figure 13 and Figure 12. Except for the case of EQ2 all other temporal datasets show increasing trend in both the number of actors as well as groups.

ACKNOWLEDGMENTS. This research is supported by NSF Grant No. IIS-1422802.

Table 1. Datasets Details

Dataset Name	Group Activity / Type	Hyper-edges	Group	Vertices	Actor	Mean Group Size	Mean Vertex Degree	Avg. No Actors Across Snapshots	Avg. No Groups Across Snapshots	Avg. Number of Active Snapshots
Pubmed	Research Publication	8683653	Group of Researches	4274803	Researcher	3.533	7.176	338814	226614	1.3
USPatent	Patent Publicaiton	1221741	Group of Researches	1330422	Researcher	2.262	2.077	118067	70979	1.3
DBLP	Research Publication	1180465	Group of Researches	1304358	Researcher	3.175	2.874	227630	119435	1.1
EQ2	Online Game Task	11240	Group of Game Players	10933	Game Player	2.91	2.992	2128	1112	18.9
Enron	Email Communication	3016	Group of Members in an Email Communication	184	Member of Organization	3.632	59.527	51	58	3.6
Apache	Code Editing	80910	Codes Editing Same Software Code	3360	Software Coder	5.08	122.340	-	-	-
IMDB	Movie Making	10174	Movie Making Group	95321	Movie Actors	22.778	2.431	-	-	-
LiveJournal	Interest Group	664414	Group of Individuals Sharing interest in Journalism	1147948	Individual	10.789	6.244	-	-	-
Friendster	Interest Group	1620991	Group of Individuals Sharing interest in Blogs	7944949	Individual	14.484	2.955	-	-	-
Youtube	Interest Group	16386	Group of Individuals Sharing interest in Online YouTube Videos	52675	Individual	7.885	2.453	-	-	-