# Finding Prominent Actors in Dynamic Affiliation Networks

Hossam Sharara
Computer Science Department
University of Maryland, College Park
hossam@cs.umd.edu

Lisa Singh
Department of Computer Science
Georgetown University, Washington DC
singh@cs.georgetown.edu

Lise Getoor
Computer Science Department
University of Maryland, College Park
getoor@cs.umd.edu

Janet Mann
Departments of Biology and Psychology
Georgetown University, Washington DC
mannj@georgetown.edu

## ABSTRACT

Most networks contain embedded communities or groups that impact the overall gathering and dissemination of ideas and information. These groups consist of important or prominent individuals who actively participate in network activities over time. In this paper, we introduce a new method for identifying actors with prominent group memberships in time-varying affiliation networks. We define a prominent actor to be one who participates in the same group regularly (stable participation) and participates across different groups consistently (diverse participation), thereby having a position of structural influence in the network. Our proposed methods for quantifying stable and diverse participation takes into consideration the underlying semantics for group participation as well as the level of impact of an actor's history on his or her current behavior. We illustrate the semantics of our measures on both synthetic and real-world data sets with varying temporal connectivity structures. We also illustrate their utility by demonstrating their complementary nature when compared to existing centrality measures.

## I MOTIVATION

Much research has focused on identifying influential individuals and opinion leaders in a social network; that is, those who have high social capital, or who help maximize the spread of information and ideas [1]. While identifying these individuals is useful for macro-level diffusion analysis, it is less useful for understanding the structural influence of individuals in embedded communities or groups in the network. In this context, we introduce the concept of prominent individuals. Based on their position in the network, they can have the greatest influence on their underlying groups. More specifically, a prominent individual is one who participates regularly within a group (stable participation) and consistently across many groups (diverse participation) compared to others in the network. These individuals are important to the network because they have access to more information than those that participate in only one or two groups and they have the potential to disseminate information since they participate consistently across many groups.

Our new method for identifying actors with prominent group membership incorporates measures that quantify two specific types of actor behavior across different groups: *stable actors*, those who participate in the same groups regularly, and *diverse* actors, those who participate across different groups consistently. Together, these measures are used to identify and rank actors with *prominent* group memberships in any time-varying network, particularly affiliation networks. An affiliation network contains two different types of nodes, one for actors and one for events, and edges between actors and the events in which actors participate [2]. In time-varying affiliation networks, an actor's participation in a particular event is associated with a specific time, indicating when the participation occurred. Many interesting social networks can be captured as affiliation networks, including organizational data describing peoples' roles on teams or in companies; and social media data describing users participation in blogs.

As an example, consider an epidemiological network where groups are based on exposure to different disease strains. Stable actors represent vulnerable individuals who are consistently and recently exposed to a certain disease. Diverse actors represent those that have repeatedly had exposure to a large number of disease groups. While each of these measures provide meaningful insight into the implications of different types of disease exposures, studying the behavior of individuals who have elongated exposure to different types of diseases is meaningful for understanding the vulnerability of the network as a whole, and the dynamics of the spread of different disease strains.

In this paper, we extend the work we proposed in [3] for quantifying user loyalty to a social group into a more general formulation for identifying prominent users based on both their stability in their corresponding groups, as well their diversity across different groups in networks. The contributions of this paper are as follows. First, we propose a new method for identifying individuals who have both stable and extended reach in a network that is tunable for networks with varying characteristics, including different semantics for network group participation and for past participation. Second, we show how our method can be used in time-varying affiliation networks; specifically, how we can use the affiliation events to define the groups that actors belong to. Third, we demonstrate the utility of the measures on both synthetic and real world data sets and compare them to existing centrality measures of influence.

This paper is organized as follows. Section II presents the related work. Section III formally defines affiliation networks and our grouping semantics. Section IV describes our stability, diversity, and prominence measures and ranking. We evaluate our methods on both real-world and synthetic data sets, and compare them to existing centrality measures in Section V. Finally, conclusions are presented in Section VI.

## II RELATED WORK

Analyzing the dynamics of social group formation and user behavior within these groups is a growing area of interdisciplinary research. The community detection literature has focused on using measures of cohesion and clustering to identify subsets of users in the network that are densely connected to each other, but less densely connected to users in other clusters. However, the majority of research conducted on community detection focuses on static networks and consider only the case when an actor can be a member of a single community [4–8].

Recently, researchers have begun to analyze the dynamics of communities over time [9–16]. Much of this research focuses on two questions: what are the communities that exist in a particular data set, and how do they change or evolve over time. The work by Tant et al. [16] is concerned with identifying a core set of communities of actors over time. Asur et al. [9] focused more on developing methods for identifying significant changes to groups over time. Their proposed approach relies on partitioning temporal snapshots into groups and analyzing the corresponding group popularity and influence indices. In addition, Sun et al. [15] introduced a parameter-free approach for discovering communities and change points in community structure without relying on temporal snapshots. Berger-Wolf and Saia [12] used social groups to partition the data, and developed algorithms for generating meta-group statistics to find a "critical group set", one whose removal leaves no visible meta-groups. Friedland and Jensen [13] introduced a method for detecting small groups of individuals sharing unusual affiliations over time.

The above methods for modeling and understanding the dynamics of the community formation and group structure can be categorized as a macro-level analysis on the underlying social network. In contrast, the approach we propose in this paper is a micro-level analysis technique that focuses on the dynamics of specific actors or individuals within different groups in the network. Our analysis of actors can be conducted using the groups identified by any of the aforementioned dynamic community detection algorithms on a single mode social network. Once the social groups or communities are established, our goal is to understand the dynamics of actors and their social relationships in the context of these predefined groups.

Backstrom et al. [11] discussed the notion of engaged users in "thriving" social groups, and discovered that the core users of the group tend to receive preferential treatment from other members. This work emphasizes the role and importance of these types of users in online social groups. Another approach by Habiba et al. [17] proposed a set of methods for identifying important actors in dynamic networks. The authors identified nodes in single mode networks that are likely to be good "spread blockers". To accomplish this, they introduced dynamic measures for density, diameter, degree, betweenness, closeness and clustering coefficient. Their measure of dynamic average degree is semantically meaningful in the context of time-varying affiliation networks. We show that it is a special case of our diversity measure when there is no discount function. While all of these approaches are important for micro-level analysis of dynamic social networks, none of them identify individuals that are structurally well positioned in the network to gather and disseminate information within different groups or communities.

# III DEFINITIONS AND BACKGROUND

In this section, we define the notions of dynamic groups and affiliation networks. We then explain a novel approach for defining groups from affiliation networks.

## 1 TEMPORAL SOCIAL GROUPS

The groups that we focus on are temporal, so an actor's participation varies with time. In addition, we allow actors to be members of multiple groups at each time step since this property is more typical in real-world networks. More formally, given a single mode graph $G(\mathcal{A}, E)$ containing a set of actor nodes $\mathcal{A} = \{a_1, a_2, a_3, \ldots, a_n\}$ and edges that connect actors, $E = \{(a_i, a_j) | a_i \text{ and } a_j \in \mathcal{A}\}$, we define a collection of groups, $\mathcal{G} = \{g_1, g_2, g_3, \ldots, g_l\}$, and a boolean relationship that describes temporal group memberships of an actor, $GroupMember(a_i, g_j, t)$. We also define a temporal social group as a subset of actors having the same group value $g_j$ at time $t$: $SocialGroup(g_j, t) = \{a_i | a_i \in \mathcal{A}, GroupMember(a_i, g_j, t)\}$.

The above construct is general and can be applied in a variety of settings depending on the semantics of the underlying analysis task. One could group actors based on the output of the dynamic community detection algorithm discussed in the previous section. Another method would be to group actors based on common attribute values or common event participation. We now describe this method in the context of affiliation networks.

## 2 EVENT-BASED GROUPING FROM AFFILIATION NETWORKS

An affiliation network is represented as a graph $G(\mathcal{A}, \mathcal{E}, \mathcal{P})$ containing a set of actor nodes $\mathcal{A} = \{a_1, a_2, a_3, \ldots, a_n\}$, a set of event nodes $\mathcal{E} = \{e_1, e_2, e_3, \ldots, e_m\}$, and a set of participation edges $\mathcal{P}$ that connect actors in $\mathcal{A}$ to events in $\mathcal{E}$. Here, $\mathcal{P} = \{(a_i, e_j) | a_i \in \mathcal{A}, e_j \in \mathcal{E}, \text{ and } a_i \text{ participates in } e_j\}$. In addition, we assume that actors and events have attributes or features associated with them. In order to emphasize the event relation's temporal component, it explicitly contains a time attribute, $E_{time}$, that can be used to associate specific time points to abstract groups of actors. An example of an affiliation network is an author publication network. In such a network, the actors are the authors, the events are the publications, and the participation relationship indicates the authors of a publication. This affiliation network is temporal because each publication event has a date of publication. We will use this author publication network as our running example.

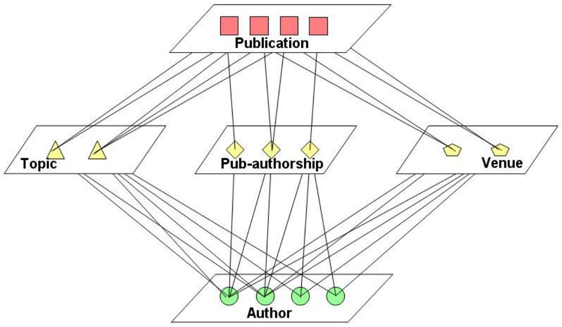


Figure 1: Grouping abstractions based on publication event attributes at time $t$

The grouping construct we propose for affiliation networks incorporates the semantics of events into the grouping definition. Specifically, we propose defining a group based on *shared event attribute values*. We consider any attribute (or combination of attributes) of the event relation to be a *grouping abstraction* that can be used to define a set of groups. Assume each attribute $E_k$ of $Event$ has a domain of values, $Domain(E_k) = \{g_1, g_2, \ldots g_p\}$. In order to construct an event-based grouping abstraction, we define the temporal group membership as follows: $GroupMember(a_i, g_j, E_{time})$ where $g_j$ is in $Domain(E_k)$, and $E_{time}$ is the time point when the group membership is valid. For simplicity, our definition focuses on the case where groups are based on a single attribute $E_k$. It is straightforward to extend this definition to consider multiple attributes.

Returning to our publication network with authors as actors and publications as events, we focus on the three attributes associated with the publication event - publication topic, publication venue, and publication authors. Each attribute is an abstraction through which actors can be grouped. Using this formulation, we have multiple ways an actor relates to other actors through a particular event. An example social group belonging to the *topic* grouping abstraction is *data mining*, e.g., actors who have published on the topic data mining. Figure 1 shows a partial lattice structure for the publication network at a particular time point $t$. We see that authors are connected to publications through similar values of three different grouping abstractions: topic, authorship and venue. This lattice structure emphasizes the connectivity between the temporal social groups and the actors and events in the network.

There are several advantages to our group formulation. First, our approach, while very simple, is surprisingly flexible. Second, actors can belong to multiple affiliation-based groups at a particular time. In other words, membership in different groups can be overlapping. Third, actors are not required to be part of an event (or group) at every time $t$.

## IV QUANTIFYING ACTOR PARTICIPATION

Formally, we are interested in the following problem: Given a dynamic affiliation network $\mathcal{G}$, identify the top-k prominent actors $P$ in the network. Intuitively, an actor that is prominent has two characteristics. First, the actor participates within a dynamic social group consistently over time and hence, is considered a *stable actor*. Second, the actor participates across many different groups consistently over time and therefore, is considered a *diverse actor*. We now define actor stability, diversity, and prominence in the remainder of this section.

### 1 ACTOR STABILITY

Different static and temporal definitions can exist for stability. We consider two in this paper: *frequency-based stability* and *consistency-based stability.*

### 1.1 FREQUENCY-BASED STABILITY

One possible definition for stability considers the number of times an actor participates in a group. Let $n_s(a_i, g_j)$ be the number of time points that actor $a_i$ participates in group $g_j$. Let $T_{max}$ be the maximum number of time points that any of the actors in the network participated in any group in $\mathcal{G}$. Then the *stability* of actor $a_i$ in a group $g_j$ is defined as the ratio between the number of time points $a_i$ participates in a particular group $g_j$ and the maximum number of time points $T_{max}$:

$$\mathcal{S}(a_i, g_j) = \frac{n_s(a_i, g_j)}{T_{max}}$$

While this definition takes into consideration the dynamics between actors and groups, it does not capture the temporal component of the actor participation.

### 1.2 CONSISTENCY-BASED STABILITY

In some domains, it is important to favor consistent and recent actor participation over irregular and outdated participations. This is especially relevant in data sets containing a large number of time periods, in which it is valuable to highlight the duration(s) for which a stable member possesses this property. For this reason, we introduce a *discount function* that serves as the mechanism for taking the temporal component of the actor's participation into account. As a result, instead of using the total number of a given actor's participations in a corresponding group when calculating stability, we can alter the effect over time depending on the discount function.

The main inputs to the discount function are the previous value of the actor's participation in group $g_j$ and the difference in time between the current time point and the previous time point where the actor's last activity was monitored, $t - t_{prev}$. The output of the function is the discounted value of the actor's participation at the current time step $t$. Examples of common discount functions are linear ($\mathcal{F}(x, y) = \frac{x}{\alpha . y}$) and exponential decay ($\mathcal{F}(x, y) = \frac{x}{e^{\alpha . y}}$). We place no constraint on the type of discount function used in the model. Any model of decay that suits the domain being studied is reasonable.

We calculate the discounted sum of the actor participation at each time point as follows:

$$\begin{aligned} \mathcal{N}_s(a_i, g_j, t) = & \quad \delta(a_i, g_j, t) \qquad\qquad\qquad t = t_0 \\ = & \quad \delta(a_i, g_j, t) \\ & \quad +(\mathcal{F}(\mathcal{N}_s(a_i, g_j, t_{prev}), t - t_{prev})) \\ & \qquad\qquad\qquad\qquad\qquad t > t_0 \end{aligned}$$

where $\mathcal{N}_s(a_i, g_j, t)$ is the discounted value of actor $a_i$'s participation in group $g_j$ up to time point $t$, $\mathcal{F}$ is the user-defined discount function, $t_0$ is the initial time point where actor $a_i$ participated in group $g_j$, $t_{prev}$ is the last time point the actor participated in before $t$, and $\delta(a_i, g_j, t)$ is a participation function that evaluates to one when actor $a_i$ participates in group $g_j$ at time $t$ or zero when actor $a_i$ does not. We then evaluate stability at the time point of interest, $t_f$, using the discounted value of the actor participation up until that point by augmenting the original stability measure as follows:

$$\mathcal{S}(a_i, g_j, t_f) = \frac{\mathcal{N}_s(a_i, g_j, t_f)}{T_{max}}$$

where $T_{max}$ is the maximum number of time points from $t_0$ until time point $t_f$ that any actor in $\mathcal{A}$ participated in any group in $\mathcal{G}$. Here, we use $T_{max}$ to normalize the scores. We could simply use the total number of time points; however, if most actors participate in a small fraction of time steps, using the maximum participation of any actor results in a wider distribution of stability values.

## 2 ACTOR DIVERSITY

Similar to stability, different static and temporal definitions can exist for diversity. Two that we consider in this paper are *frequency-based diversity* and *consistency-based diversity.*

### 2.1 FREQUENCY-BASED DIVERSITY

One possible definition for actor diversity considers the number of groups in which an actor participates. Let $n_d(a_i)$ represent the number of groups that actor $a_i$ participates in over all time points and let $G_{max}$ be the total number of groups with at least one actor participation in the network, where $G_{max} \leq |\mathcal{G}|$. Then the *diversity* of actor $a_i$ is defined as the number of groups $a_i$ actually participates in over the number of groups $a_i$ can participate in:

$$\mathcal{D}(a_i) = \frac{n_d(a_i)}{G_{max}}$$

### 2.2 CONSISTENCY-BASED DIVERSITY

Similar to stability, we are interested in favoring recent and consistent diversity. Therefore, we will also use a discount function for the actor's diversity. We first calculate the discounted sum of actor participations at each time point:

$$\begin{aligned} \mathcal{N}_d(a_i,t) = &\quad n_d(a_i,t) & t = t_0 \\ = &\quad n_d(a_i,t) & \\ &\quad +(\mathcal{F}(\mathcal{N}_d(a_i,t_{prev}),t-t_{prev})) & \\ & & t > t_0 \end{aligned}$$

where $\mathcal{N}_d(a_i,t)$ is the discounted value of actor $a_i$'s number of group participations up to time point $t$, $\mathcal{F}$ is the user-defined discount function, and $t_{prev}$ is the last time point the actor participated in prior to $t$. Then, the diversity at the time point of interest, $t_f$, is calculated using the discounted value of the actor's number of group participations up to $t_f$ divided by the number of groups in the network times the number of time points the actor participated in.

$$\mathcal{D}(a_i,t_f) = \frac{\mathcal{N}_d(a_i,t_f)}{G_{max} \times T_{max}}$$

where $T_{max}$ is the maximum number of time points that any of the actors in *Actor* participated in any group in $\mathcal{G}$ until time point $t_f$.

## 3 ACTOR PROMINENCE

Once we have the stability for all actors in their corresponding groups, as well as their overall diversity, we can use these measures to determine the set of prominent actors in the affiliation network.

**DEFINITION 1.** *A prominent actor $P$ has both a high stability $\mathcal{S}$* ***within*** *groups in $\mathcal{G}$ and a high diversity $\mathcal{D}$* ***across*** *groups in $\mathcal{G}$ over time.*

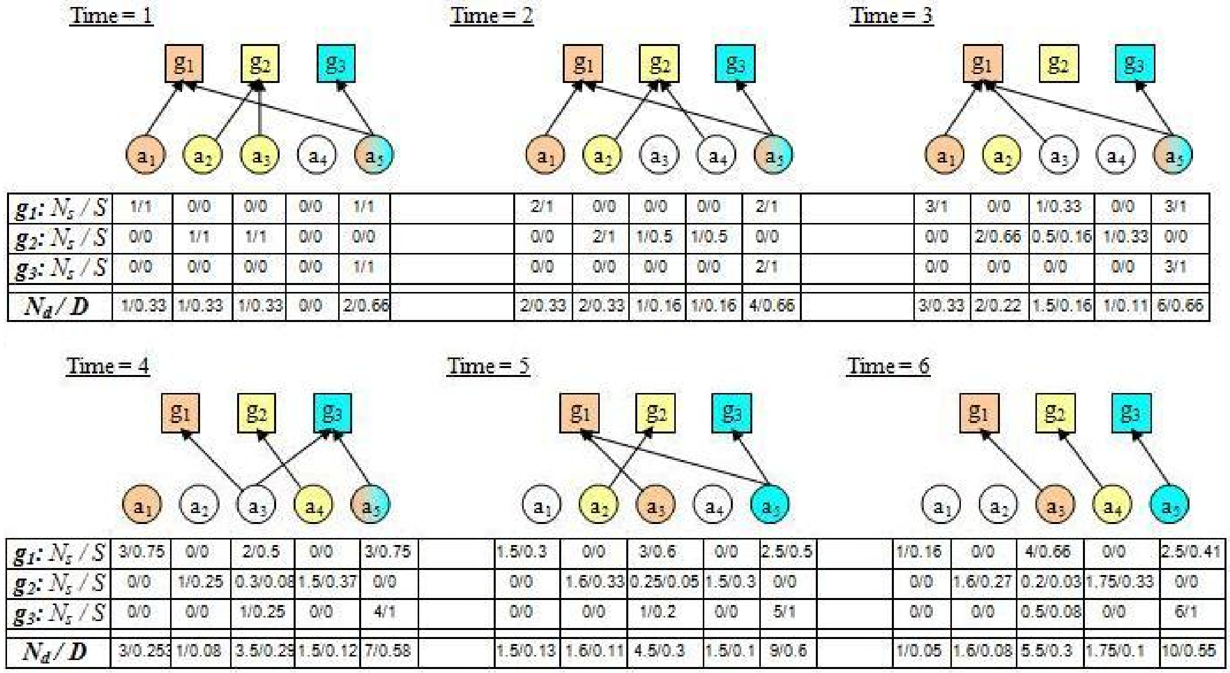
We define $\mathcal{SA}_k(g_j,t)$ to be the top-k stable actors for group $g_j$ at time point $t$ and $\mathcal{SA}_k(G,t) = \bigcup_{g_j \in \mathcal{G}} \mathcal{SA}_k(g_j,t)$. Similarly, we define $\mathcal{DA}_k(t)$ to be the top-k diverse actors at time point $t$. Then prominence $P(t)$ is calculated as follows:

1. Calculate $\mathcal{D}(a_i,t)$ and $\mathcal{S}(a_i,g_j,t)$ for all $a_i$ and $g_j$.
2. Determine $\mathcal{SA}_k(G,t)$ and $\mathcal{DA}_k(t)$.
3. Intersect top-k stability and diversity sets to find
prominent actors $P_k(t) = \mathcal{SA}_k(G,t)) \cap \mathcal{DA}_k(t))$.

This final set will contain the actors who possess both high stability and high diversity measures. Notice that it is possible for a particular data set to contain no prominent actors. For example, there may be affiliation networks that contain stable members that are not diverse. In such cases, the intersection will yield an empty set of prominent actors. To avoid elevating non-prominent actors in data sets containing low diversity and stability values for all the actors, we can include a minimum threshold so that only actors above the minimum thresholds for stability and diversity are candidates for prominence.

## 4 EXAMPLE CALCULATIONS

In order to further illustrate these measures, consider the simple example in Figure 2 containing 5 actors participating in 3 groups over 6 time points, with a linear discount function. First, we consider stability.



| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $g_1$: $N_s$ / $S$ | 1/1 | 0/0 | 0/0 | 0/0 | 1/1 | | 2/1 | 0/0 | 0/0 | 0/0 | 2/1 | | 3/1 | 0/0 | 1/0.33 | 0/0 | 3/1 |
| $g_2$: $N_s$ / $S$ | 0/0 | 1/1 | 1/1 | 0/0 | 0/0 | | 0/0 | 2/1 | 1/0.5 | 1/0.5 | 0/0 | | 0/0 | 2/0.66 | 0.5/0.16 | 1/0.33 | 0/0 |
| $g_3$: $N_s$ / $S$ | 0/0 | 0/0 | 0/0 | 0/0 | 1/1 | | 0/0 | 0/0 | 0/0 | 0/0 | 2/1 | | 0/0 | 0/0 | 0/0 | 0/0 | 3/1 |
| $N_d$ / $D$ | 1/0.33 | 1/0.33 | 1/0.33 | 0/0 | 2/0.66 | | 2/0.33 | 2/0.33 | 1/0.16 | 1/0.16 | 4/0.66 | | 3/0.33 | 2/0.22 | 1.5/0.16 | 1/0.11 | 6/0.66 |

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $g_1$: $N_s$ / $S$ | 3/0.75 | 0/0 | 2/0.5 | 0/0 | 3/0.75 | | 1.5/0.3 | 0/0 | 3/0.6 | 0/0 | 2.5/0.5 | | 1/0.16 | 0/0 | 4/0.66 | 0/0 | 2.5/0.41 |
| $g_2$: $N_s$ / $S$ | 0/0 | 1/0.25 | 0.3/0.08 | 1.5/0.37 | 0/0 | | 0/0 | 1.6/0.33 | 0.25/0.05 | 1.5/0.3 | 0/0 | | 0/0 | 1.6/0.27 | 0.2/0.03 | 1.75/0.33 | 0/0 |
| $g_3$: $N_s$ / $S$ | 0/0 | 0/0 | 1/0.25 | 0/0 | 4/1 | | 0/0 | 0/0 | 1/0.2 | 0/0 | 5/1 | | 0/0 | 0/0 | 0.5/0.08 | 0/0 | 6/1 |
| $N_d$ / $D$ | 3/0.253 | 1/0.08 | 3.5/0.29 | 1.5/0.12 | 7/0.58 | | 1.5/0.13 | 1.6/0.11 | 4.5/0.3 | 1.5/0.1 | 9/0.6 | | 1/0.05 | 1.6/0.08 | 5.5/0.3 | 1.75/0.1 | 10/0.55 |

Figure 2: Example for calculating actor stability and diversity.

We assign a color to each group in the figure. Group $g_1$ is (orange), $g_2$ is (yellow), and $g_3$ is (blue). The most stable actor in each group at every time point is assigned the same color of the group. For example, at time point 1, actors $a_1$ and $a_5$ are the most stable actors in group $g_1$. Actors $a_2$ and $a_3$ are most stable in group $g_2$. Actor $a_5$ is also most stable in group $g_3$. Notice that actor $a_5$ is stable in two groups at time point 1 and therefore, is two colors. The first three rows under each time period show each actor's discounted sum for stability and the stability for groups $g_1$, $g_2$, and $g_3$, respectively. The last row shows each actor's discounted sum for diversity and the diversity for each actor.

Let's first consider the stability of the actors. By examining the evolution of actor stability, we see that until the third time point, the stable members of each group do not change because of the persistent participation of those actors in their corresponding groups. At the third time point, the stability of actor $a_2$ drops. By the fourth time point, the stability score of actor $a_2$ in group $g_2$ is lower than that of actor $a_4$ since actor $a_2$ has not participated in group $g_2$ for two time points and actor $a_4$, a previous participant of group $g_2$, begins participating in group $g_2$ again.

At the fifth time point, the stable members of both groups $g_1$ and $g_2$ change. Actor $a_3$, who consistently participated in group $g_1$ during the last 3 time points, has a higher stability score than both actor $a_1$, who has stopped participating in the group since the third time point, and actor $a_5$, who did not participate in the group in the previous time point. Thus, actor $a_3$ becomes the most stable actor in group $g_1$. As for group $g_2$, since actor $a_2$ returns, $a_2$'s stability score increases over that of actor $a_4$ who missed the participation at this step. Finally, at the final time point, actor $a_4$ participates in group $g_2$ and becomes more stable than actor $a_2$.

At the final time point, we can see how the temporal aspect of our proposed measures impacts the results. For group $g_1$, the most stable member is actor $a_3$ who participated in the group consistently and recently over the last 4 time points. The second most stable member is actor $a_5$ who also participated in the group at 4 time points, but in earlier time points than that of $a_3$. The least stable member of the group is actor $a_1$ who participated in only 3 earlier time points. As for group $g_2$, we find that the scores of both actors $a_2$ and $a_4$ are very close. Actor $a_4$ is considered more stable because he or she participates more recently and more consistently than actor $a_2$. Finally, for group $g_3$, actor $a_5$ is the most stable actor. Since he or she participated in group $g_3$ at every time point, his or her stability score is 1.

As for diversity, we can see that actor $a_5$ is also the most consistently diverse actor over time, having the highest diversity score at the final time point. Actor $a_5$ regularly appears in two of the three groups. In contrast, actors $a_1$ and $a_2$ have the lowest diversity scores since they appear in only a single group throughout the example. In order to realize the importance of including the temporal discount factor, we notice that if the discount factor is not included, the diversity of both actors $a_3$ and $a_5$ would be the same even though actor $a_5$ is consistently diverse for 4 time points and actor $a_3$ participated in two groups only once.

Suppose we set $k = 2$ to determine the $k$ prominent actors. The two most stable actors are $a_5$ and $a_3$. These two actors also have the highest diversity scores and are therefore, both prominent actors.

## V EXPERIMENTAL RESULTS

We begin by analyzing our proposed measures, stability, diversity, and prominence, on three affiliation networks: a scientific publication network, a senate bill sponsorship network, and a dolphin social network. We analyze the distribution of values for each data set and illustrate meaningful characteristics of the actors in the networks. We then compare stability and diversity to well known centrality measures and show that these measures capture a different dynamic than existing measures. To show how our measures perform on a broader range of actor behaviors, we generate a set of synthetic data sets that contain varying probabilities for generating different types of actors and use our measures to see if they adequately identify stable and diverse actors. Finally, we show how the different temporal discount functions can be used, illustrating the value of using dynamic measures rather than static ones.

### 1 DATA SETS

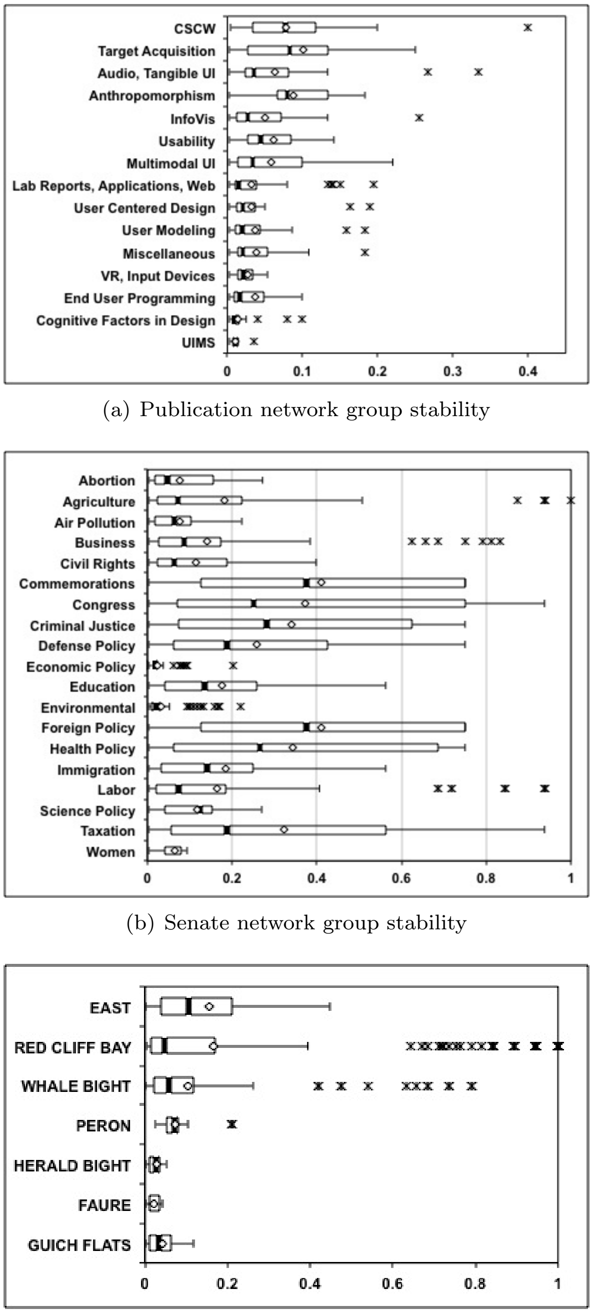
**Scientific publication network:** This network is based on publications in the ACM Computer-Human Interaction (ACMCHI) conference from 1982 until 2004. Similar to our running example, this data set describes an author/publication affiliation network. It was extracted from the ACM Digital Library and contains 4,073 publications and 6,358 authors. There are 12,727 participation relationships (edges) between authors and publications. Since we are interested in the temporal dynamics of the actors, single actor participations are removed as a preprocessing step for all the data sets. We grouped publications using the *topic* attribute. There are 15 values for this attribute.

**Senate bill sponsorship network:** This network is based on data collected about senators and the bills they sponsor [18]. The data contains each senator's demographic information and the bills each senator sponsored or co-sponsored from 1993 through February 2008. Each bill has a date and topics associated with it. We group the bills using their general topic. After removing the senators that do not sponsor a bill, the bills that do not have a topic, and preprocessing the data, our analysis uses 181 senators, 28,372 bills, and 188,040 participation relationships spanning 100 general topics. While we used all the groups for our analysis, due to space limitations we illustrate the results using only a subset of the 100 topics.

**Dolphin behavioral network:** This network is based on a data set accumulated over the last 25 years on a population of wild bottle-nose dolphins in Shark Bay, Australia. The dolphin population has been monitored annually since 1984 by members of the Shark Bay Dolphin Research Project. They have collected 13,400 observation surveys of dolphin groups. Each observation of a group of dolphins represents a 'snapshot' of associations and behaviors. In this affiliation network, dolphins are defined as actors and surveys as events. Dolphins observed in a survey constitutes the participation relationship. Our analysis includes 560 dolphins, 10,731 surveys, and 36,404 relationships between dolphins and surveys. We group survey observations together by the location (latitude-longitude) of the survey. Seven different predetermined areas of approximately equal size (75 sq km.) were used as groups for this data set.

### 2 MEASURING STABILITY

The results of measuring the stability of actors to different groups in each network are summarized in Figures 3(a) - 3(c). Here the x-axis represents the stability value and the y-axis contains the group names. Except where otherwise noted, we use a linear discount function with $\alpha = 1$ for all the results reported. As can be seen from the figures, because the semantics and evolution of each network are different, the overall actor stability varies across the data sets with the average stability being lowest for the publication data set and highest for the senate data set. The low average stability of the authors to publication topics



(a) Publication network group stability

(b) Senate network group stability

(c) Dolphin network group stability

Figure 3: Stability of actors across various groups

in the scientific publication data set results because most of the authors do not publish in this venue every year on the same topic. This may lead one to believe that these authors are diverse. As we will see later, while there are some diverse authors, the majority are not. In fact, the majority of authors do not consistently stay a member of a single group and do not consistently publish across groups. In this data set, the low stability scores is an indication that very few people in the network are well positioned to have a strong, continual impact on authors in these different groups.

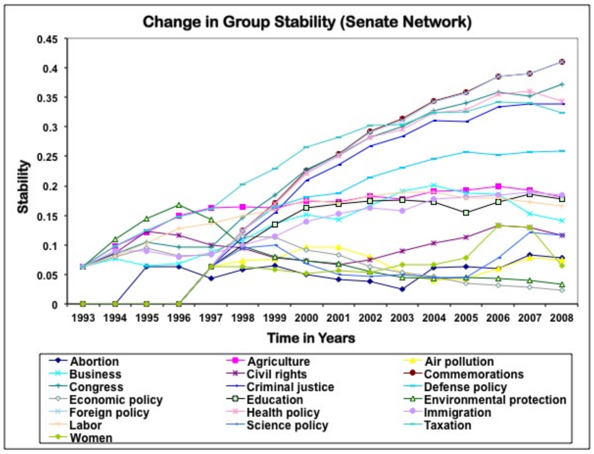


Figure 4: Group stability changes in senate network

Figure 3(b) shows the results on the senate bill sponsorship network. Here we notice that the average stability of senators in different groups is much higher than that of the publication network with some having a score above 0.8. This means that senators are regularly sponsoring bills of a certain type. This is particularly true for bills sponsored within certain topics, e.g., commemorations, foreign policy, taxation. However, other topics, including environmental policy and women, have significantly less stable membership (average stability less than 0.1). A more detailed temporal comparison is illustrated in Figure 4. It shows the changing dynamics of the average actor stability in different groups over time. Initially, they are relatively similar and relatively low. However, since the late 1990's, the average stability values have increased rapidly for six groups including defense, commemorations, business, and criminal justice, while the remaining average stability values have been relatively consistent, with only a slight increase or decrease, each year. The increase in stability of senators in these areas is consistent with historical events, e.g. wars in Afghanistan and Iraq.

Figure 3(c) shows the results on the dolphin network. Here we see that dolphin stability is more variable than the previous data sets. There are three locations with more stable membership than the others. The average stability is highest for the 'East' location, but the most stable dolphins in the data set are

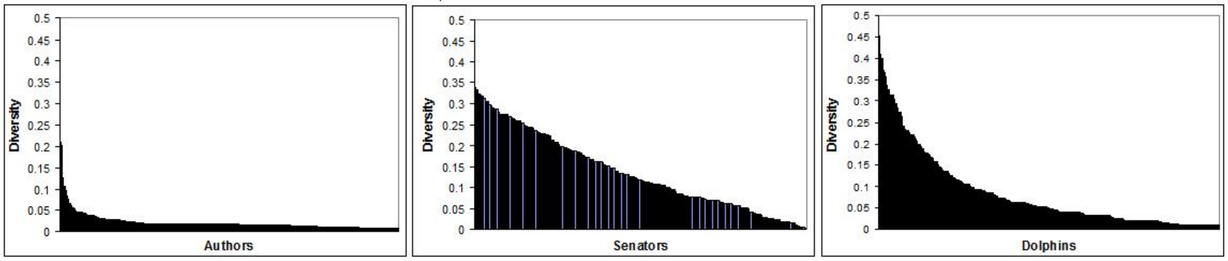


Figure 5: Actor diversity

in 'Red Cliff Bay' and 'Whale Bight'. While some of this may be explained by heavier sampling in certain regions, biologists believe this is likely to be a result of habitat structure in the region [19, 20]. For example, 'East', which has the highest average stability, is mostly deep channels bisected by shallow sea grass banks. Dolphins with high stability in the 'East' have certain foraging specializations (channel foragers or sea grass bed foragers). Since many dolphins spend a large amount of time foraging, a high stability in regions where specialized foraging is necessary is consistent with biologists' interpretation of dolphin behavior. Dynamic measures like stability provide observational scientists with a tool for measuring and comparing social variability throughout an animal's life history.

## 3 MEASURING DIVERSITY

In Figure 5, we compute the diversity distribution among the actors of each network. To make the figure easier to read, we sorted the diversity values for each data set from highest to lowest along the x-axis. The figure shows that the average diversity is highest for senators, while the range of diversity values is widest for the dolphins. The diversity of actors in the scientific publication network is very low (average $< 0.05$). This is an indication that authors are not publishing consistently across topics. The diversity values for the senator sponsorship network may seem low since they are all below 0.5 and intuitively, we expect senators to sponsor bills across a range of topics. However, this is not surprising because there are 100 different bill topics, and the number of topics is part of the denominator of the diversity equation. Finally, the range in dolphin diversity is much higher than the other two data sets. Again, this is consistent with biologists' interpretation of dolphin behavior. While many dolphins settle in some areas (bights or bays), others spend more time in adjacent bays at specific stages in their life history (e.g., juvenile period or adulthood), thereby increasing their diversity score with respect to location.

## 4 PROMINENT ACTORS

Recall that prominent actors are structurally well positioned in the network to both gather new information and ideas from different groups (diversity), as well as disseminate them to members of groups they actively participate in (stability). In order to find the prominent actors, we apply the method discussed in Section IV using ($k$=10). We highlight some interesting findings. First, none of the data sets have 10 prominent actors. In other words, few actors in the data sets are both stable and diverse. The dolphin data set, which has the highest stability and diversity scores, returns the fewest prominent actors (4). The senator network has the largest number of prominent actors (8), and the publication data set is in the middle (6).

Focusing on the senator data set, the following actors are considered prominent: Sen. Jeff Bingaman, Sen. Barbara Boxer, Sen. Diane Fienstein, Sen. Edward Kennedy, Sen. John Kerry, Sen. Patrick Leahy, Sen. Joseph Lieberman, and Sen. Patricia Murray - all well-known Democratic senators. To gain further insight we determined whether or not the prominent actors are stable in the same groups. Figure 6 shows

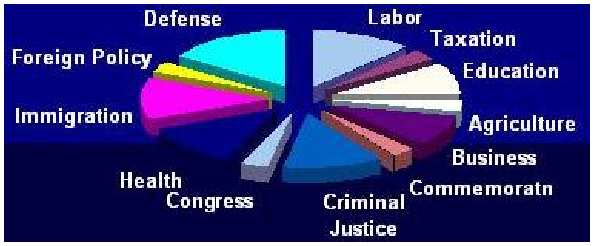
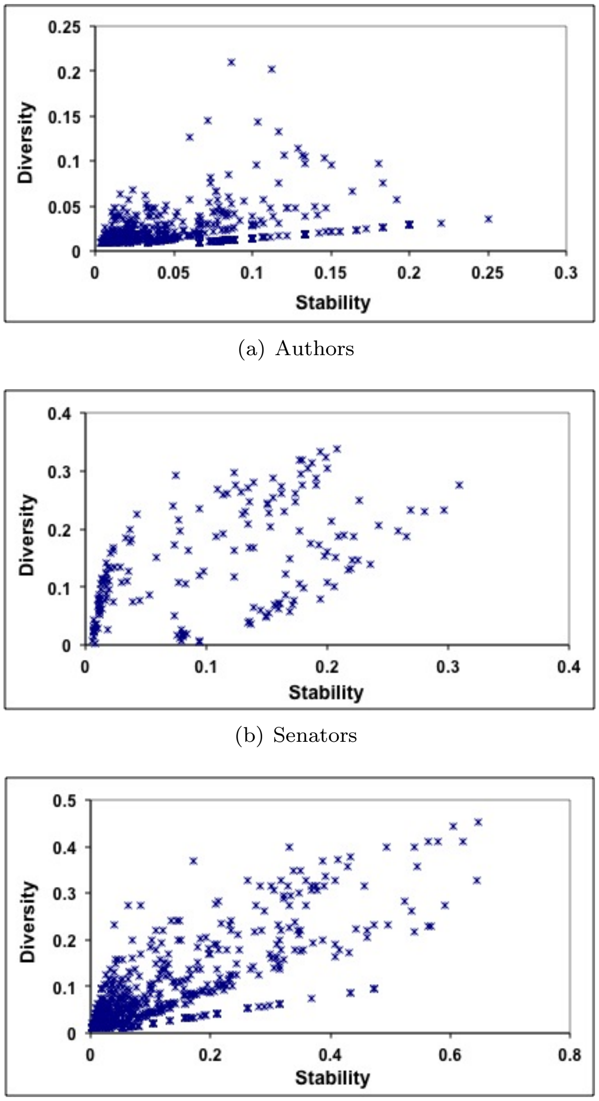


Figure 6: Stable group membership of prominent senators



(a) Authors

(b) Senators

(c) Dolphins

Figure 7: Stability vs Diversity

the groups in which these senators are considered stable. The size of each pie slice represents the number of prominent actors in the topic group. While there is definite overlap (5 out of 8 are stable in defense), there are clear differences as well (Sen. Patty Murray from Washington is stable in Agriculture).

We have similar findings for the publication data set. We also further analyze the prominent actors in the publication network by checking the DBLP listing of publications for each prominent author. We find that for all the authors, 31% to 53% of their total publications are in ACM CHI. Thus, this conference represents a very important venue in their research portfolio. Finally, prominent actors in the dolphin network all have high stability and diversity in the same location groups. Scientists who monitor the dolphins know these dolphins to be highly sociable, i.e. since the 1980s, their rate of contact with other dolphins is high. These dolphins are sighted regularly in many different locations and are rarely sighted alone. This is consistent with the definition of prominence.

In order to better understand the relationships between stability and diversity of all the actors in the data sets, we plotted stability vs. diversity in Figure 7. The variations across data sets is evident. For the publication network, high stability generally correlates with low diversity. The senator network is much more varied, while in the dolphin network highly stable actors also tend to be more diverse. These figures reconfirm that prominence is not a common characteristic for individuals in different networks and prominent individuals hold a unique position in the network that provides them an opportunity to be very influential in groups within the network.

## 5 COMPARISON WITH CENTRALITY MEASURES

A natural direction is to understand how stability and diversity compare to existing centrality measures. Do they capture the same information, or do they provide additional insight? We begin by comparing stability to the most common centrality measures. In order to do this, we generated the underlying single-mode, co-membership network for actors participating in a certain affiliation group, and computed various centrality measures on the generated networks. We show the results for the publication data in Table 1 using the 'Information Visualization' topic as a sample affiliation group. The table shows the measure values, followed by the ranking of the actor for each measure. For example, Benjamin Bederson ranked as the author with highest betweenness and eigenvector centrality. However, by examining the publications pattern we note that they are neither consistent across time nor numerous, and the same is true for Robert Spence who was ranked first according to the closeness centrality. On the other hand, the time-consistent, recent and numerous publications of the most stable author, namely Stuart Card, illustrates exactly what our proposed stability measure captures that the other centrality measures missed. A similar experiment was performed using the original affiliation network to compare our proposed diversity measure with the same centrality

| **Actor** | Stability | Closeness | Betweenness | Eigenvector |
|---|---|---|---|---|
| | *Raw Value(Rank)* | | | |
| | **Participation** [*Publication Year*] | | | |
| Stuart K. Card | 0.255(1) | 0.305(85) | 0.445(6) | 0.003(9) |
| | 1987, 1991, 1993, 1994, 1995, 1996, 1998, 2000, 2001, 2003, 2004 | | | |
| Robert Spence | 0.038(49) | 1(1) | 0.006(50) | 0(88) |
| | 1993, 1996, 1999, 2001 | | | |
| Benjamin B. Bederson | 0.117(18) | 0.321(82) | 1(1) | 0.276(1) |
| | 1995, 1999, 2000, 2003 | | | |

Table 1: Stability vs. Centrality

| **Actor** | Diversity | Closeness | Betweenness | Eigenvector |
|---|---|---|---|---|
| | *Raw Value(Rank)* | | | |
| | **Participation** [*Year(Group Count)*] | | | |
| Allison Druin | 0.209(1) | 0.395(10) | 0.013() | 0.002() |
| | 1994(2), 1995(1), 1996(2), 1997(1), 1998(2), 1999(4), 2000(2), 2001(2), 2002(4), 2003(1), 2004(1) | | | |
| Ben Shneiderman | 0.127(6) | 0.446(1) | 0.3(2) | 0.0027(1) |
| | 1982(1), 1987(1), 1991(2), 1992(3), 1993(1), 1994(3), 1995(2), 1996(1), 1998(5), 1999(2), 2001(1), 2002(4) | | | |
| Brad A. Myers | 0.106(9) | 0.428(3) | 0.306(1) | 0.0025(6) |
| | 1985(1), 1987(1), 1990(1), 1991(3), 1992(1), 1993(4), 1994(3), 1995(3), 1996(3), 1998(1), 2000(2), 2002(3) | | | |

Table 2: Diversity vs. Centrality

measures. Although the results reported in Table 2 have more similarity, the effects of recency and consistency over time capture a dynamic that the other centrality measures miss.

## 6 EVALUATION ON SYNTHETIC DATA

To illustrate that our measure effectively captures the semantics we expect, we developed a synthetic data generator that allows us control the behavior of actors within and across affiliation groups. Our data generator uses an evolutionary approach to emulate the dynamics of a time-varying affiliation network. We allow the actors in the network to exhibit 4 different behaviors: normal, stable, diverse and persistent.

The generator starts by creating an initial network containing $n_g$ affiliation groups and $n_a$ actors using a predefined percentage of each type of actor. When an actor is created, it samples its lifetime from the distribution $D_L$, and picks its type according to the input probabilities. Then, the network is allowed to evolve through the simulation time $T$, with specific evolution rules according to each actor's type. A normal actor samples the number of affiliation links to be created at this time from the distribution $D_N$, then picks a random set of groups corresponding to this number. The other types of actors sample their number of participations from the (higher) distribution $D_F$, where a stable actor creates its participation in the one group it was affiliated with in a prior time point, a diverse actor chooses a random set of groups to participate in, and a prominent actor establishes a set of groups that it continues participating in over time. Actors are generated throughout the simulation according to the same probabilities to keep the average number of active nodes and the distribution of the actor types fixed over time.

The data generator is very flexible, allowing users to specify the following parameters: the total period of simulation $T$, the average number of active actors at any time point $n_a$, the number of affiliation groups $n_g$, the probability of generating a stable actor $p_s$, a diverse actor $p_d$, or a prominent actor $p_p$, a distribution of the actors' lifetime $D_L$, and two distributions

| $p_s$ | $p_d$ | $p_p$ | $F_1(diversity)$ | $F_1(stability)$ | $F_1(prominence)$ |
|---|---|---|---|---|---|
| 0.2 | 0.2 | 0.04 | 0.992 | 0.976 | 0.931 |
| 0.1 | 0.1 | 0.01 | 0.99 | 0.974 | 0.985 |
| 0.01 | 0.01 | 0.0001 | 0.945 | 0.989 | 1 |

Table 3: $F_1$ measures on synthetic data

of actors' participations per time step $D_N$ and $D_F$, where the first is used for normal actors (with relatively lower participation) and the second is used for the other types of actors (with higher participation).

To better understand how well our measures capture the different actor behaviors, we generated a number of synthetic networks with different distributions of actor types. The synthetic networks used for these experiments were obtained by setting the parameters of the generator as follows: ($T = 25$ time points, $n_a = 1000$ actors, $n_g = 25$ groups, $D_L$ $N(12, 3)$, $D_N$ $N(0, 1)$, and $D_F$ $N(3, 1)$). It is important to note that the distributions for the user participations should follow the same distribution, but with a lower mean for the 'normal' actors. Doing so is necessary to generate the required behavior for diverse and prominent actors. We carried out the experiments by varying the probabilities for generating different types of actors, and then using our measures on the different networks to see if we could correctly identify the stable, diverse and prominent actors at the different time points.

For calculating stability, diversity and prominence, we used a linear discount function with a value of ($\alpha = 1$), and a threshold equal to the corresponding percentage of active actors that should be present at the network at the end of the simulation. Each experiment for a certain set of parameters was carried out 100 times and the average F-1 measure for stability (as a group average), diversity, and prominence are reported in Table 3. The table shows that for networks with different distributions of actor behaviors, our measures accurately identify stable, diverse and prominent actors, with an F-1 measure of over 0.9 in all cases.

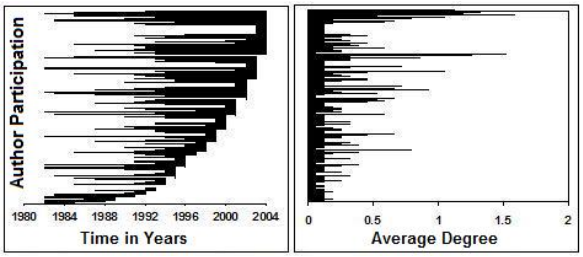


Figure 8: Exponential discount ($\mathcal{F}(x, y) = \frac{x}{e^y}$)

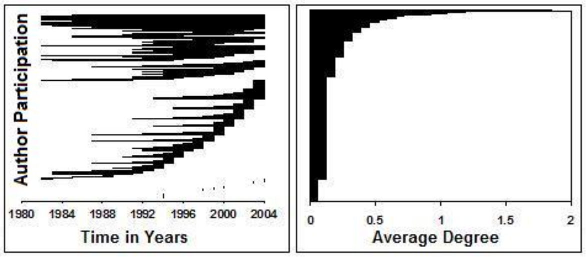


Figure 9: Identity discount ($\mathcal{F}(x, y) = x$)

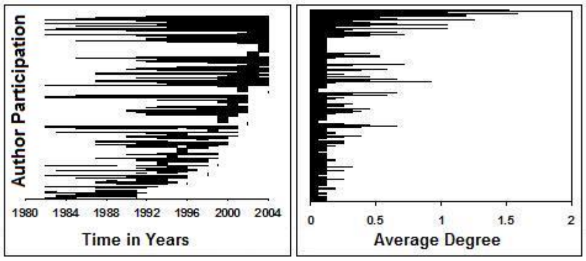


Figure 10: Linear discount ($\mathcal{F}(x, y) = \frac{x}{y}$)

## 7 COMPARISON OF DISCOUNT FUNCTIONS

In order to demonstrate the flexibility introduced in our model through the discount function, we illustrate the results of using different temporal discount functions to calculate the diversity of authors in the publication data set. We consider three different models for decay - an exponential, the identity (i.e., none) and a linear model. The results are illustrated in Figures 8, 9 and 10 respectively. The left part of the figures shows the authors participation over time. The right side shows the average degree of those actors. The actors in both graphs are sorted by increasing diversity.

The exponential decay discount function favors recency, as illustrated Figure 8. Actors that participate through the entire time period, including most recently, have the highest diversity. The diversity of those actors that appear in the same number of time periods differs, depending upon recency. The top diverse actors are determined based on the recency of their last group participation since the effect of their older participations is diminished exponentially. This model is useful when we need to capture the recent behavior of actors with only up-to-date participations, with limited influence from their past behavior.

At the other extreme, the model could use the identify function and effectively do no temporal discounting. In this case, the order of occurrence of actor participations in different groups is ignored, and the top actors are determined only by the number of participations in their corresponding groups. Using this model is equivalent to using the dynamic average degree proposed by Habiba et al. [17]. Analyzing Figure 9, we see that actors having the same number of participations at different time periods have the same diversity value. This model is appropriate when those analyzing the network are more concerned with the magnitude of the actors' group participations over time.

Finally, the linear model attempts to account for both recency and frequency in determining top actors. As we can see in Figure 10, the time graph shows that the model does still favor recency, but it also captures the magnitude of group participations (as shown in the right-hand graph). Such a model can be used when we need to account for recent actor behavior, older behavior, and consistent behavior.

## VI CONCLUSIONS AND FUTURE WORK

In this paper, we introduce the concepts of stable, diverse and prominent actors in a network and exhibit methods for identifying them in the case of dynamic affiliation networks. Because these networks are more nuanced than traditional static social networks, the measures are more complex, and capture both the temporal aspects of the networks and the variety of ways of defining groups within an affiliation network. We illustrate the utility of our measures of stability and diversity on several real-world networks, compare them to other well known measures of centrality, and show how they highlight important subtleties that are not captured traditionally. We also show how our proposed measures can be used to accurately capture prominent actors that are persistent within groups and diverse across groups on different synthetic data sets. Finally, we analyze prominent actors in different domains and highlight the importance of capturing both stability and diversity.

One direction for future work is to adapt the proposed method for characterizing the dynamic evolution of different ties between actors. Another interesting direction is to investigate the exact role that these prominent actors play in the dissemination of information across a network, as well as the general dynamic underlying network formation in the presence of such actors. It would also be interesting to integrate these measures into a learning algorithm that predicts future network dynamic. Last but not least, analyzing the variations in the proposed measures could lead to valuable insights about the specific nature of different network types.

## VII ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their valuable feedback and the members of the Shark Bay Dolphin Research Project (SBDRP). This work was supported in part by ONR Grant #10230702 and NSF Grants #0941487 and #0746930.

A summary version based on this paper appears in ASE Social Informatics 2012. This paper was originally submitted to that conference and accepted by the ASE Social Informatics 2012 PC as a journal paper.

### References

[1] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network", in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2003.

[2] S. Wasserman and K. Faust, *Social network analysis: methods and applications*, Cambridge University Press, Cambridge, 1994.

[3] H. Sharara, L. Singh, L. Getoor, and J. Mann, "Understanding actor loyalty to event-based groups in affiliation networks", *Social Network Analysis and Mining*, vol. 1, pp. 115–126, 2011.

[4] D. Cai, Z. Shao, X. He, X. Yan, and J. Han, "Community mining from multi-relational networks", in *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, 2005.

[5] G. Flake, S. Lawrence, C. Giles, and F. Coetzee, “Self-organization and identification of web communities”, *Computer*, vol. 35, no. 3, pp. 66–71, 2002.

[6] J. Hopcroft, O. Khan, B. Kulis, and B. Selman, “Natural communities in large linked networks”, in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2003.

[7] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney, “Statistical properties of community structure in large social and information networks”, in *Proceedings of the 17th International World Wide Web Conference (WWW)*, 2008.

[8] M. E. J. Newman, “Detecting community structure in networks”, *The European Physical Journal B*, vol. 38, pp. 321–330, 2004.

[9] S. Asur, S. Parthasarathy, and D. Ucar, “An event-based framework for characterizing the evolutionary behavior of interaction graphs”, in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2007.

[10] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, “Group formation in large social networks: membership, growth, and evolution”, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

[11] Lars Backstrom, Ravi Kumar, Cameron Marlow, Jasmine Novak, and Andrew Tomkins, “Preferential behavior in online groups”, in *Proceedings of the first ACM International Conference on. Web Search and Data Mining (WSDM)*, 2008.

[12] T. Berger-Wolf and J. Saia, “A framework for analysis of dynamic social networks”, in *Proceeding of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

[13] L. Friedland and D. Jensen, “Finding tribes: identifying close-knit individuals from employment patterns”, in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2007.

[14] T. A.B. Snijders, “Models for longitudinal network data”, *In P. Carrington, J. Scott, and S. Wasserman (Eds.), Models and methods in social network analysis. New York: Cambridge University Press*, p. Chapter 11, 2005.

[15] J. Sun, C. Faloutsos, S. Papadimitriou, and P. Yu, “Graphscope: parameter-free mining of large time-evolving graphs”, in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2007.

[16] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe, “A framework for community identification in dynamic social networks”, in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2007.

[17] Habiba, T. Y. Berger-Wolf, Y. Yu, and J. Saia, “Finding spread blockers in dynamic networks”, in *the 2nd SNA-KDD Workshop on Social Network Mining and Analysis*, 2008.

[18] Govtrack, “Senate bill sponsorship data”, website: www.govtrack.us, 2008.

[19] B. L. Sargeant, J. Mann, P. Berggren, and M. Krützen, “Specialization and development of beach hunting, a rare foraging behavior, by wild indian ocean bottlenose dolphins”, *Canadian Journal of Zoology*, vol. 83, pp. 1400–1410, 2005.

[20] J. Mann and B. Sargeant, “Like mother, like calf: The ontogeny of foraging traditions in wild indian ocean bottlenose dolphins”, *In D. Fragaszy and S. Perry, The Biology of Traditions: Models and Evidence. Cambridge University Press*, pp. 236–266, 2003.