# Stability vs. Diversity: Understanding the Dynamics of Actors in Time-varying Affiliation Networks

Hossam Sharara*, Lisa Singh†, Lise Getoor* and Janet Mann‡
*Computer Science Department, University of Maryland, College Park
†Department of Computer Science, Georgetown University, Washington DC
‡Departments of Biology and Psychology, Georgetown University, Washington DC

***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 real-world data sets with varying temporal connectivity structures.**

## 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. 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 to denote those individuals who, through their position in the network, can have the greatest influence on its underlying groups. More specifically, a *prominent* individual is one that 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.

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 had elongated exposure to different types of diseases is crucial in understanding the vulnerability of the network as a whole, and the dynamics of spread of different disease strains.

In this paper, we extend the work we proposed in [1] 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 three real world data sets. An extended version of this paper is presented in [2].

## II. Related Work

Recently, researchers have begun to analyze the dynamics of communities over time [3]–[10]. 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.

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.

Habiba et al. [11] 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.

## 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.

### A. 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 a 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.

### B. 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.

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.

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.

### A. Actor Stability

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

*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.

*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 the case of data sets with large number of time periods, in which it is valuable to highlight the duration(s) for which a stable member possesses this property. To this extent, 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 used.

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 which function is used. 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:

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

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.

### *B. 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*.

*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) 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 $t_f$.

### *C. Actor Prominence*

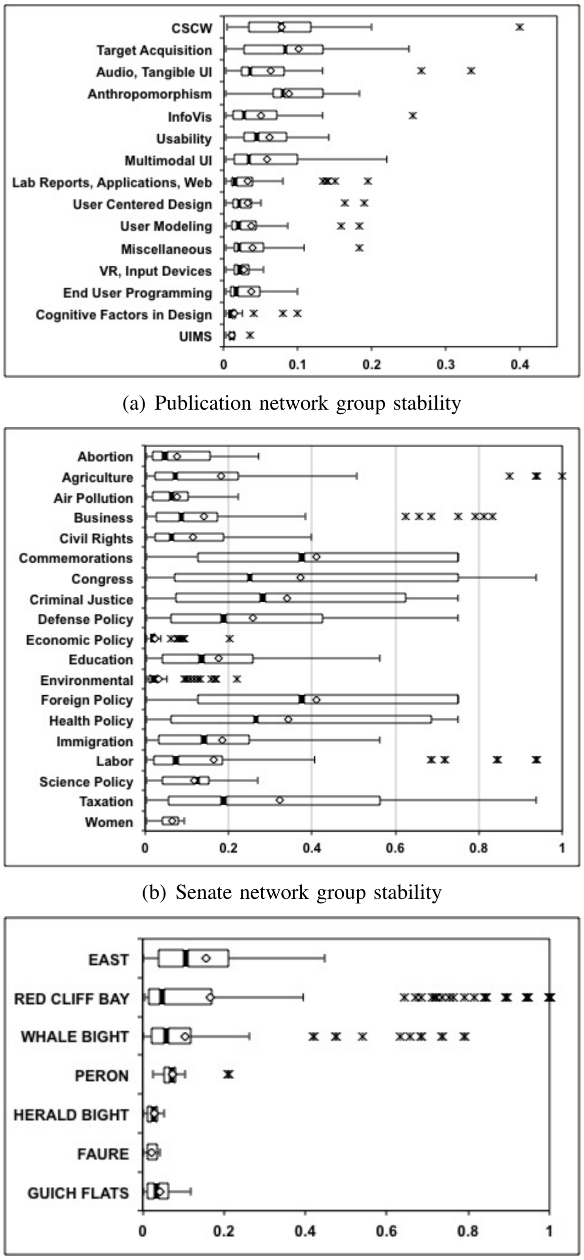
Once we have the stability for all actors in their corresponding groups, as well as their overall diversity, we can use these measures to determine the set of prominent actors in the affiliation network.

*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



(a) Publication network group stability

(b) Senate network group stability

(c) Dolphin network group stability

Fig. 1. Stability of actors across various groups

threshold so that only actors above the minimum thresholds for stability and diversity are candidates for prominence.

## 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. A more extensive set of experiments are presented in the extended version of this paper [2].

### A. 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 [12]. The data contains each senator's demographic information and the bills each senator sponsored or co-sponsored from 1993 through February 2008. There are 1896 senators, 28,949 bills, and 191,097 sponsorship/co-sponsorship relationships between the senators and the bills in this affiliation network. 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.

### B. Measuring Stability

The results of measuring the stability of actors to different groups in each network are summarized in Figures 1(a) - 1(c). Here the x-axis represents the stability value and the y-axis contains the group names. Except where noted otherwise, 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

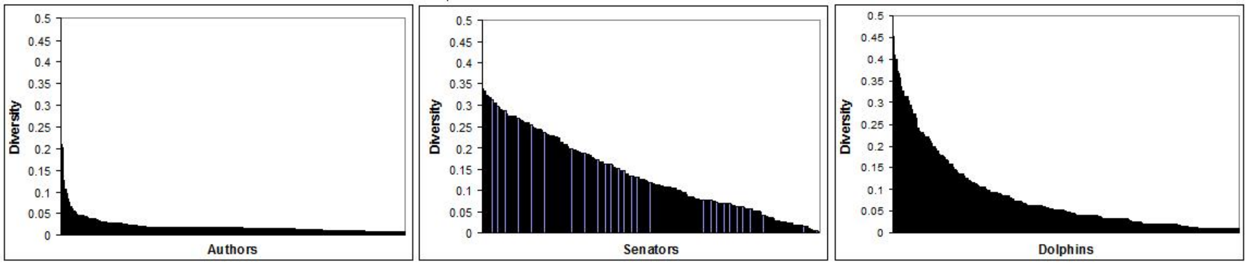


Fig. 2. Actor diversity

the authors to publication topics 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 1(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).

Figure 1(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 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 [13]. 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.

### C. Measuring Diversity

In Figure 2, 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.

### D. 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-C 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. We have similar findings for the publication data set. 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.

## VI. Conclusions

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 then 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.

## 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.

## References

[1] 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.

[2] ——, "Finding prominent actors in dynamic affiliation networks," *To appear: Human Journal*, 2012.

[3] S. Asur, S. Parthasarathy, and D. Ucar, "An event-based framework for characterizing the evolutionary behavior of interaction graphs," in *KDD*, 2007, pp. 913–921.

[4] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: membership, growth, and evolution," in *KDD*, 2006.

[5] L. Backstrom, R. Kumar, C. Marlow, J. Novak, and A. Tomkins, "Preferential behavior in online groups," in *WSDM*, 2008.

[6] T. Berger-Wolf and J. Saia, "A framework for analysis of dynamic social networks," in *KDD*, 2006.

[7] L. Friedland and D. Jensen, "Finding tribes: identifying close-knit individuals from employment patterns," in *KDD*, 2007.

[8] T. A. 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.

[9] J. Sun, C. Faloutsos, S. Papadimitriou, and P. Yu, "Graphscope: parameter-free mining of large time-evolving graphs," in *KDD*. New York, NY, USA: ACM, 2007, pp. 687–696.

[10] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe, "A framework for community identification in dynamic social networks," in *KDD*, 2007.

[11] Habiba, T. Y. Berger-Wolf, Y. Yu, and J. Saia, "Finding spread blockers in dynamic networks," in *SNA-KDD*, 2008.

[12] Govtrack, "Senate bill sponsorship data," website: www.govtrack.us, 2008.

[13] 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.