

An Event-based Framework for Characterizing the Evolutionary Behavior of Interaction Graphs

SITARAM ASUR
SRINIVASAN PARTHASARATHY ¹
and
DUYGU UCAR
Ohio State University

Interaction graphs are ubiquitous in many fields such as bioinformatics, sociology and physical sciences. There have been many studies in the literature targeted at studying and mining these graphs. However, almost all of them have studied these graphs from a static point of view. The study of the evolution of these graphs over time can provide tremendous insight on the behavior of entities, communities and the flow of information among them. In this work, we present an event-based characterization of critical behavioral patterns for temporally varying interaction graphs. We use non-overlapping snapshots of interaction graphs and develop a framework for capturing and identifying interesting events from them. We use these events to characterize complex behavioral patterns of individuals and communities over time. We show how semantic information can be incorporated to reason about community-behavior events. We also demonstrate the application of behavioral patterns for the purposes of modeling evolution, link prediction and influence maximization. Finally, we present a diffusion model for evolving networks, based on our framework.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms: Algorithms, Measurement

Additional Key Words and Phrases: Dynamic Interaction networks, Evolutionary analysis, Diffusion of innovations

1. INTRODUCTION

Many social and biological systems can be represented as complex interaction networks where nodes represent entities (i.e. individuals, proteins) and edges mimic the interactions among them. These interaction networks arise from a wide variety of scientific domains like computer science, physics, biology and sociology. Online communities such as Flickr, MySpace and Orkut, e-mail networks, co-authorship networks and WWW networks are examples of interesting interaction networks. The study of these complex interaction networks can provide insight into their structure, properties and behavior.

Early research in this area [Girvan and Newman 2002; Newman 2006; Barabasi and Bonabeau 2003; Barabasi et al. 2002] has primarily focused on the static properties of these networks, neglecting the fact that most real-world interaction networks

¹Contact Author: Srinivasan Parthasarathy, Department of Computer Science, Ohio State University, 395 Dreese Labs, 2015 Neil Ave, Columbus, OH 43210.
Email: srini@cse.ohio-state.edu

are dynamic in nature. In reality, many of these networks constantly evolve over time, with the addition and deletion of edges and nodes representing changes in the interactions among the modeled entities. Recently, there has been some interest in studying dynamic graphs [Leskovec et al. 2005; Backstrom et al. 2006; Chakrabarti et al. 2006; Falkowski et al. 2006]. Identifying the portions of the network that are changing, characterizing the type of change, predicting future events (e.g. link prediction), and developing generic models for evolving networks are challenges that need to be addressed. For instance, the rapid growth of online communities has dictated the need for analyzing large amounts of temporal data to reveal community structure, dynamics and evolution.

Interaction networks are often modular in nature. The interactions existing between nodes can be used to group them into clusters or communities. For instance, in a social network, these clusters represent people with similar contact patterns or interests. The problem of identifying these clusters or communities from static graphs has been extensively studied [Girvan and Newman 2002; Clauset et al. 2004; Flake et al. 2002] over the past decade. However, in the case of evolving graphs, the clusters are typically not static. Instead they constantly change over time, as the network evolves. We believe that studying the evolution of these clusters, in particular their formation, transitions and dissolution, can be extremely useful for effectively characterizing the corresponding changes to the network over time.

Another important aspect is the behavior of the nodes of the network. Nodes of an evolving interaction network represent entities whose interaction patterns change over time. The movement of nodes, their behavior and influence over other nodes, can help make inferences regarding future interactions as well as predicting changes to communities in the network. For instance, in a social network, if a person is very sociable, the chances of him/her interacting with new people and joining new groups is very high. In the case of a collaboration network, if a person is known to collaborate frequently with different people, then the chance of a new collaboration involving this person is high. The influence exerted by a node can be studied in terms of its effect on other nodes. If several other people join a community when a particular individual does, it indicates a high degree of positive influence for that person.

Another important factor influencing behavior and evolution is the semantic nature of the interactions itself. For instance, in a co-authorship network, two authors are connected if they publish a paper together. The topic or the subject area of the paper will definitely influence future collaborations for each of these authors. If there are different authors working on similar topics, the chances of them collaborating in the future is higher than two authors working on unrelated areas. We wish to examine the influence of the semantics of the interaction on future interactions.

The study of diffusion or flow of information in an evolving network is important for social science research, viral marketing applications and epidemiology. For instance, pandemic viruses pose a severe threat to society due to their potential to spread rapidly and cause tremendous widespread illnesses and deaths. In viral marketing, the goal is to propagate an idea or innovation through an interaction network. Analysis of the evolution of interactions in a social network and the identification of influential nodes can be used to devise effective containment poli-

cies for pandemic disease spread in the case of epidemiological studies, as well as ‘word-of-mouth’ advertising in marketing.

In this paper, we provide an event-based framework for characterizing the evolution of interaction networks. We begin by converting an evolving graph into static snapshot graphs at different time points. We obtain clusters at each of these snapshots independently. Next, we characterize the transformations of these clusters by defining and identifying certain critical events. We define efficient incremental algorithms involving bit-matrix computations for this purpose. We use these critical events to compute and reason about novel behavior-oriented measures, which offer new and interesting insights for the characterization of dynamic behavior of interaction graphs. We also develop semantic information measures that can be used to analyze and reason about community behavior. We illustrate our framework on three different evolving networks - the DBLP co-authorship network, Wikipedia and a clinical trials patient network. In each case, the behavioral patterns that we discover using our framework help us make useful inferences about cluster evolution and link prediction. Finally, we use the behavioral measures to detail a diffusion model for evolving networks and demonstrate their application for the task of influence maximization.

In short, the key contributions of this work are

- (1) The identification of key critical events that occur in evolving interaction networks.
- (2) Efficient incremental algorithms for the discovery of these critical events
- (3) Novel behavioral measures for stability, sociability, influence and popularity that can be computed incrementally over time
- (4) A diffusion model for evolving networks based on our framework
- (5) Application of the events and behavioral measures on two real datasets for modeling evolution, predicting behavior and trends (link prediction) and influence maximization.

2. RELATED WORK

There has been enormous interest in mining interaction graphs for interesting patterns in various domains. However, the majority of these studies [Girvan and Newman 2002; Newman 2006; Barabasi and Bonabeau 2003; Barabasi et al. 2002; Clauset et al. 2004; Flake et al. 2002] have focused on mining static graphs to identify community structures, patterns and novel information. Recently, the dynamic behavior of clusters and communities have attracted the interest of several groups.

Leskovec et al. [2005] studied the evolution of graphs based on various topological properties, such as the degree distribution and small-world properties of large networks. They empirically showed that these networks become denser over time, with the densification following a power-law pattern. They also found that the effective diameter of these networks decreases as the network grows. They proposed a graph generation model, called *Forest Fire* model, based on their findings on the evolutionary behaviors of graphs. In later work, Leskovec et al. [2008] conducted an extensive analysis of community structure on 70 social and information networks. To aid their analysis, they made use of a goodness measure called Conductance,

and a Network-Profile plot, which plots the quality of communities against their sizes. They concluded empirically that the Conductance scores of the best possible sets of nodes deteriorated as the sets got larger.

Backstrom et al. [2006] studied formation of groups and the ways they grow and evolve over time. They proposed using features of communities and individuals, applying decision-tree techniques to estimate the probability of an individual joining a community, as well as identifying communities that are likely to grow.

Chakrabarti et al. [2006] proposed evolutionary settings for two widely-used clustering algorithms (k-means and agglomerative hierarchical clustering). They defined evolutionary clustering as the task of incrementally obtaining high-quality clusters for a set of objects while also maintaining similarity with clusters identified in previous timestamps. To obtain the clusters for a particular snapshot, they also used history information to obtain a clustering consistent with earlier snapshots. Chi et al. [2007] have developed an evolutionary version of the spectral clustering algorithm making use of temporal smoothness, to obtain consistent clusters. They have shown how their method can provide optimal solutions to relaxed versions of the evolutionary kmeans algorithm.

Falkowski et al. [2006] analyzed the evolution of communities that are stable or fluctuating based on subgroups. They have examined overlapping snapshots of interaction graphs and applied standard statistical measures to identify persistent subgroups. Tantipathananandh et al. [2007] have developed a framework for detecting dynamic community structure in evolving graphs. They make use of dynamic programming and heuristics to optimize the structure discovered. Although these researchers have analyzed interaction graphs, their focus is different from ours. Our focus is on identifying key events and behavioral patterns that can characterize, model and predict future behavioral trends. We are interested in discovering changes that occur over time, rather than merely identifying community structure.

Palla et al. [2007] have performed dynamic analysis on evolving collaboration and phone-call networks. They have shown empirically that the lifetime of groups or communities in these networks depends on the dynamic behavior of these groups, with large groups that alter their behavior persisting longer than others. On the other hand, small groups were found to persist longer if their membership remained unchanged. This work was concurrent to our original technical report [Asur et al. 2007], and has a different direction from our work. While their study has been primarily on analyzing lifetime of communities and its relation with size, ours is more broader and concerns all possible changes occurring over time for both communities and nodes. We have presented a framework for characterizing these changes and using them to construct measures for quantifying dynamic behavior. Hopcroft et al. [2004] have looked at tracking communities in interaction networks over time. Their goal is to identify natural stable communities that have low degrees of change and track their evolution. On the other hand, we are considering all communities and nodes and using patterns in the changes that occur to identify various types of evolutionary behavior such as stability, sociability, popularity and influence. Recently, Ferlez et al. [2008] have presented an algorithm, *TimeFall* which focuses on the analysis of the evolution of networks, across time through the evolution of its communities. However, the difference in their work from ours, is that they treat

the problem as a compression problem and aim at finding natural patterns (cuts, communities) to generate a parameter-free solution.

The seminal paper by Samtaney et al. [1994] described an approach for extracting coherent regions from 2-dimensional and 3-dimensional scalar and vector fields for tracking purposes. To study the evolution of these regions over time, they presented certain evolutionary events for objects. Yang et al. [2005] used events to mine spatial and temporal datasets. In their work, they first identified Spatio-temporal episodes (SOAPS) of a single snapshot and used events to identify behaviors of these SOAPS across snapshots.

The use of Semantic similarity based on ontologies has been studied many times in the past [Lin 1998; Ganesan et al. 2003]. Resnik [1999] suggested a novel way to evaluate semantic similarity in an ontology based on notion of information content. This notion of semantic similarity has been successfully applied on various taxonomies. Richardson et al. [1994], developed a semantic similarity measure using the WordNet knowledge base for the task of information retrieval. To measure semantic similarity, they employed a combination of a conceptual distance-based approach and the information-based approach proposed by Resnik. Lord et al. [2003], applied semantic similarity based on information content for the task of information retrieval using the Gene Ontology (GO)², as the knowledge base. Since GO is a hierarchical ontology, they use the information content of the lowest common subsumer of two genes to measure the semantic similarity between them. In our work, semantic similarity concepts are used to quantify similarity for clusters over time.

3. DATASETS

We employ three different datasets in this work.

3.1 DBLP co-authorship network:

The DBLP bibliography maintains information on more than 800000 computer science publications. We used the DBLP data to generate a co-authorship network representing authors publishing in several important conferences in the field of databases, data mining and AI. We chose all papers over a 10 year period (1997-2006) that appeared in 28 key conferences spanning mainly these three areas. The conferences we considered are - (*PKDD, ACL, UAI, NIPS, KR, KDD, ICML, ICCV, IJCAI, CVPR, AAAI, ER, COOPIS, SSDBM, DOOD, SSD, FODO, DAS-FAA, DEXA, ICDM, IDEAS, CIKM, EDBT, ICDT, ICDE, VLDB, PODS, SIG-MOD*). We converted this data into a co-authorship graph, where each author is represented as a node and an edge between two authors corresponds to a joint publication by these two authors. The graph spanning 10 years contained 23136 nodes and 54989 edges. We chose the snapshot interval to be a year, resulting in 10 consecutive snapshot graphs. These graphs are then clustered and analyzed to identify critical events and patterns. We believe that studying the evolution of the DBLP dataset can afford information about the nature of collaborations and the factors that influence future collaborations between authors.

²<http://db.yeastgenome.org/cgi-bin/GO/goTermFinder>

3.2 Wikipedia Dataset

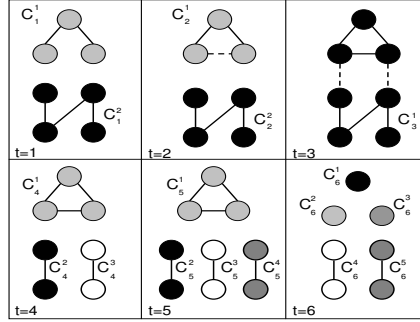
The Wikipedia online encyclopedia is a large collection of webpages providing comprehensive information concerning various topics. The dataset we employ represents the Wikipedia revision history. It consists of a set of webpages as well as links among them. It comprises of the editing history from January 2001 to December 2005. The temporal information for the creation and deletion of nodes (pages) and edges (links) are also provided. Since our primary goal for using this data is to perform semantic analysis, we make use of a category hierarchy, which we have obtained from Gabrilovich [Gabrilovich and Markovitch 2007]. The Wikipedia revision history dataset consisted of around 1.6 million webpages (nodes). We chose all webpages for which category information was available³. The resulting dataset consisted of 779005 nodes (webpages), 32.5 M edges and 78664 categories. We constructed snapshots of 3 month intervals, and considered the first 6 snapshots for our analysis. Each snapshot consisted only of nodes and edges (links) that existed during that particular interval.

3.3 Clinical Trials Data:

In clinical trials, pharmaceutical companies test a new drug for efficacy and toxicity - efficacy to evaluate its effectiveness in curing or controlling the disease in question and toxicity to determine if the drug is safe for consumption and with minimal side effects. Releasing a drug that turns out to be toxic can cost companies billions of dollars and more importantly lead to loss in life. In this paper we use a dataset obtained from a major pharmaceutical company, consisting of both healthy people as well as patients suffering from certain diseases (diabetes and hepatic impairment). As part of the study, they were given either a placebo (a formulation that includes only the inactive ingredients) or the drug under study. Liver toxicity information can be obtained from eight serum analytes, often referred to in the literature as the liver panel. This dataset consists of patients on the placebo and the drug in the ratio 40:60. The initial snapshot of this data is composed of the measurements of the analytes obtained before patients were treated with the drug or the placebo. The subsequent snapshots correspond to measurements taken every week. The data thus consists of 7 snapshots spanning a 6 week period since the beginning of the treatment. We transformed the data for each snapshot into a graph, based on the correlations that exist between the analyte values of patients. If there exists a high correlation (greater than a threshold T_{corr}) in the analyte values between two patients, the two patients have an edge between them in the snapshot graph⁴. Note that, in pharmaceutical research, it has been suggested that it is beneficial to model correlations among patients, as opposed to considering each patient in isolation [Otey et al. 2006]. The reason for this is as follows. If each patient is considered separately, we are limited to only intrinsic information, whereas by modeling patients as a graph, one can utilize intrinsic as well as extrinsic properties.

³We also included some pages that did not have category information but had strong links to the chosen pages

⁴We examined the distribution of correlations and using advice from domain-experts, picked a T_{corr} value of 0.7 for our experiments


 Fig. 1. Temporal Snapshots at time $t=1$ to 6

4. CRITICAL EVENTS

In this section, we introduce and afford a formal definition to certain critical events that occur in evolving graphs. Some of the critical events described in this section are inspired by a similar notion described by Samtaney et al. [1994], in the context of tracking and visualizing features. They have also been used for tracking spatial objects [Yang et al. 2005].

The events that we define are primarily between two consecutive timestamps but it is possible to coalesce events from contiguous timestamps by analyzing the meta-data collected from the event mining framework. We proceed to use these events in the later sections to define more complex behavior. We distribute the critical events which graphs can undergo into two categories - events involving communities and events involving individuals.

Figure 1 displays a set of snapshots of the network which will be used as a running example in this section. At time $t = 1, 2$ clusters are discovered (shown in different colors).

4.1 Events involving communities:

We define 5 basic events which clusters can undergo between any two consecutive time intervals or steps. Let S_i and S_{i+1} be snapshots of S at two consecutive time intervals with C_i and C_{i+1} denoting the set of clusters respectively. The five proposed events are:

1) Continue: A cluster C_{i+1}^j is marked as continuation of C_i^k if V_{i+1}^j is the **same** as V_i^k . We do not impose the constraint that the edge sets should be the same.

$$\text{Continue}(C_i^k, C_{i+1}^j) = 1 \text{ iff } V_i^k = V_{i+1}^j$$

The main motivation behind this is that if certain nodes are always part of the same cluster, any information supplied to one node will eventually reach the others. Therefore, as long as the vertex set remains same, the information flow is not hindered. The addition and deletion of edges merely indicates the strength between the nodes. An example of a continue event is shown at $t=2$ in Figure 1. Note that an extra interaction appears between the nodes in Cluster C_2^1 but the clusters do not change.

2) κ -Merge: Two different clusters C_i^k and C_i^l are marked as merged if there exists a cluster in the next timestamp that contains at least $\kappa\%$ of the nodes belonging to these two clusters. The essential condition for a merge is :

$Merge(C_i^k, C_i^l, \kappa) = 1$ iff $\exists C_{i+1}^j$ such that

$$\frac{|(V_i^k \cup V_i^l) \cap V_{i+1}^j|}{Max(|V_i^k \cup V_i^l|, |V_{i+1}^j|)} > \kappa\% \quad (1)$$

and $|V_i^k \cap V_{i+1}^j| > \frac{|C_i^k|}{2}$ and $|V_i^l \cap V_{i+1}^j| > \frac{|C_i^l|}{2}$. This condition will only hold if there exist edges between V_i^k and V_i^l in timestamp $i + 1$. Intuitively, it implies that new interactions have been created between nodes which previously were part of different clusters. This caused $\kappa\%$ ⁵ of nodes in the two original clusters to join the new cluster. Note that, in an ideal or complete merge, with $\kappa = 100$, all nodes in the two original clusters are found in the same cluster in the next timestamp. The two original clusters are completely lost in this scenario. We use the second term in the denominator in the above equation to differentiate cases where the merged cluster is very large and in which case the identities of the original clusters are lost. These do not qualify as *Merge* events.

Figure 1 shows an example of a complete merge event at $t=3$. The dotted lines represent the newly created edges. All the nodes now belong to a single cluster (C_3^1).

3) κ -Split: A single cluster C_i^j is marked as split if $\kappa\%$ of nodes from this cluster are present in 2 different clusters in the next timestamp. The essential condition is that:

$Split(C_i^j, \kappa) = 1$ iff $\exists C_{i+1}^k, C_{i+1}^l$ such that

$$\frac{|(V_{i+1}^k \cup V_{i+1}^l) \cap V_i^j|}{Max(|V_{i+1}^k \cup V_{i+1}^l|, |V_i^j|)} > \kappa\% \quad (2)$$

and $|V_{i+1}^k \cap V_i^j| > \frac{|C_{i+1}^k|}{2}$, $|V_{i+1}^l \cap V_i^j| > \frac{|C_{i+1}^l|}{2}$.

Intuitively, a split signifies that the interactions between certain nodes are broken and not carried over to the current timestamp, causing the nodes to part ways and join different clusters. Also note that a broken edge, by itself, does not necessarily indicate a split event, as there may be other interactions existing between vertices in the cluster (similar to the notion of k -connectivity). Time $t=4$ in Figure 1 shows a split event when a cluster gets completely split into three smaller clusters.

4) Form: A new cluster C_{i+1}^k is said to have been formed if none of the nodes in the cluster were grouped together at the previous time interval i.e. no 2 nodes in V_{i+1}^k existed in the same cluster at time period i .

$Form(C_{i+1}^k) = 1$ iff \exists no C_i^j such that $V_{i+1}^k \cap V_i^j > 1$

Intuitively, a form indicates the creation of a new community or new collaboration. Figure 1 at time $t=5$ shows a form event when two new nodes appear and a new

⁵We used κ values of 30 and 50 in our experiments.

cluster is formed.

5) Dissolve: A single cluster C_i^k is said to have dissolved if none of the vertices in the cluster are in the same cluster in the next timestamp i.e. no two entities in the original cluster have an interaction between them in the current time interval.

$$Dissolve(C_i^k) = 1 \text{ iff } \exists \text{ no } C_{i+1}^j \text{ such that } V_i^k \cap V_{i+1}^j > 1$$

Intuitively, a dissolve indicates the lack of contact or interactions between a group of nodes in a particular time period. This might signify the breakup of a community or a workgroup. Figure 1 at time $t=6$ shows a dissolve event when there are no longer interactions between the three nodes in Cluster C_5^1 resulting in a breakup of the cluster into 3 clusters - C_6^1 , C_6^2 and C_6^3 .

4.2 Events involving individuals:

We wish to analyze not only the evolution of communities but the influence of the behavior of individuals on communities. In this regard, we introduce four basic transformations involving individuals over snapshots.

1) Appear: A node is said to appear when it occurs in C_i^j but was not present in any cluster in the earlier timestamp.

$$Appear(v, i) = 1 \text{ iff } v \notin V_{i-1} \text{ and } v \in V_i$$

This simple event indicates the introduction of a person (new or returning) to a network. In Figure 1, at time $t = 5$ two new nodes appear in the network.

2) Disappear: A node is said to disappear when it was found in a cluster C_{i-1}^j but is not present in any cluster in the timestamp i .

$$Disappear(v, i) = 1 \text{ iff } v \in V_{i-1} \text{ and } v \notin V_i$$

This indicates the departure of a person from a network. In Figure 1, at time $t = 6$ two nodes of cluster C_5^4 disappear from the network.

3) Join: A node is said to join cluster C_i^j if it exists in the cluster at timestamp i . This may be due to an *Appear* event or due to a leave event from a different cluster. Note that in case, the cluster C_i^j must be sufficiently similar to a cluster C_{i-1}^k .

$$Join(v, C_i^j) = 1 \text{ iff } \exists C_{i-1}^k \text{ and } C_{i-1}^k \text{ such that } C_{i-1}^k \cap C_i^j > \frac{|C_{i-1}^k|}{2} \text{ and } v \notin V_{i-1}^k \text{ and } v \in V_i^j$$

The cluster similarity condition ensures that C_i^j is not a newly formed cluster. This condition differentiates a *Join* event from a *Form* event. Nodes forming a new cluster will not be considered to be *Join* events since there will be no cluster C_{i-1}^k in the previous timestamp with similarity $> \frac{|C_{i-1}^k|}{2}$ with the newly formed cluster.

4) Leave: A node is said to leave cluster C_{i-1}^k if it no longer is present in a cluster with most of the nodes in V_{i-1}^k . A node that leaves a cluster may leave the network as a *Disappear* event or may join a different cluster. In a collaboration network, a

Leave event might correspond to a student graduating and leaving a group.

$Leave(v, C_i^j) = 1$ iff $\exists C_i^k$ and C_{i-1}^k such that $C_{i-1}^k \cap C_i^j > \frac{|C_{i-1}^k|}{2}$ and $v \in V_{i-1}^k$ and $v \notin V_i^j$

The similarity constraint between the two clusters is used to maintain cluster correspondence. Note that if the original cluster dissolves, the nodes in the cluster are not said to participate in a *Leave* event. This is due to the fact that there will no longer be a cluster with similarity $> \frac{|C_{i-1}^k|}{2}$ with the dissolved cluster C_{i-1}^k .

4.3 Algorithms for Event Extraction:

We leverage the use of efficient bit matrix operations to compute the events between snapshots. First, for each temporal snapshot, we construct a binary $k_i \times n$ matrix T_i where k_i is the number of clusters at timestamp i and n is the number of nodes. We then compare the matrices of successive snapshots to find events between them⁶. Let $T_i(x, :)$ and $T_i(:, y)$ correspond to the x^{th} row and y^{th} column vector of matrix T_i respectively. To compute all the events between two snapshots, we perform a set of binary operations (AND and OR) on the corresponding matrices. The linear operations performed to identify each event are presented below. Let $|x|_1$ represent the L^1 -norm of a binary vector x .

$$|x|_1 = \sum_{i=1}^{|x|} x_i \quad (3)$$

We can compute the events as :

$Dissolve(T_i, T_{i+1}) = \{x | 1 \leq x \leq k_i, \arg \max_{1 \leq y \leq k_{i+1}} (|AND(T_i(x, :), T_{i+1}(y, :))|_1 \leq 1)\}$

$Form(T_i, T_{i+1}) = Dissolve(T_{i+1}, T_i)$

$Merge(T_i, T_{i+1}, \kappa) = \{ \langle x, y, z \rangle | 1 \leq x \leq k_i, 1 \leq y \leq k_i, x \neq y, 1 \leq z \leq k_{i+1}, |AND(OR(T_i(x, :), T_i(y, :)), T_{i+1}(z, :))|_1 \geq (\kappa \times Max(|OR(T_i(x, :), T_i(y, :))|, |T_{i+1}(z, :)|)), |AND(T_i(x, :), T_{i+1}(z, :))|_1 \geq \frac{|T_i(x, :)|_1}{2}, |AND(T_i(y, :), T_{i+1}(z, :))|_1 \geq \frac{|T_i(y, :)|_1}{2} \}$

$Split(T_i, T_{i+1}, \kappa) = Merge(T_{i+1}, T_i, \kappa)$

$Continue(T_i, T_{i+1}) = \{ \langle x, y \rangle | 1 \leq x \leq k_i, 1 \leq y \leq k_{i+1}, OR(T_i(x, :), T_{i+1}(y, :)) == AND(T_i(x, :), T_{i+1}(y, :)) \}$

$Appear(T_i, T_{i+1}) = \{v | 1 \leq v \leq |V|, |T_i(:, v)|_1 == 0, |T_{i+1}(:, v)|_1 == 1\}$

$Disappear(T_i, T_{i+1}) = \{v | 1 \leq v \leq |V|, |T_i(:, v)|_1 == 1, |T_{i+1}(:, v)|_1 == 0\}$

$Join(T_i, T_{i+1}) = \{ \langle y, v \rangle | 1 \leq y \leq k_{i+1}, 1 \leq v \leq |V|, T_{i+1}(y, v) == 1, \exists x, 1 \leq x \leq k_i \text{ s.t. } |AND(T_i(x, :), T_{i+1}(y, :))|_1 > \frac{|T_i(x, :)|_1}{2}, T_i(x, v) == 0 \}$

$Leave(T_i, T_{i+1}) = \{ \langle x, v \rangle | 1 \leq x \leq k_i, 1 \leq v \leq |V|, T_i(x, v) == 1, \exists y, 1 \leq y \leq k_{i+1} \text{ s.t. } |AND(T_i(x, :), T_{i+1}(y, :))|_1 > \frac{|T_i(x, :)|_1}{2}, \}$

⁶If the number of nodes changes between the timestamps, we will increase the length of the matrices to reflect the largest of the two

Time stamps	DBLP		Wikipedia	
	Active Nodes	Time (secs)	Active Nodes	Time (secs)
1-2	0.23	0.088	0.03	0.12
2-3	0.25	0.094	0.07	0.5
3-4	0.24	0.087	0.13	1.7
4-5	0.26	0.099	0.19	4.5
5-6	0.27	0.091	0.22	11.15
6-7	0.29	0.096		
7-8	0.34	0.12		
8-9	0.41	0.14		
9-10	0.40	0.14		

Table I. Timing Results for Event Detection for DBLP and Wikipedia.

$T_{i+1}(y, v) == 0\}$

$T_i(x, y)$ represents the value in the x^{th} row and y^{th} column of T_i . The construction of the matrices and the operations to find the events are all linear in time complexity($O(n)$), assuming that $k_i \ll n$ and $k_{i+1} \ll n$. The advantage of using the bit matrix operations is that they enable us to leverage GPU and multi-core architectures quite efficiently. Note that, the whole event detection process lends itself easily to parallelization.

Also, it is important to note that for the cluster membership matrix, it is not necessary to consider all n nodes. To compute events between two time points i and $i + 1$ we only need to consider nodes that are active in either of these two intervals. This greatly reduces the size of the cluster matrix describe above, since the number of columns would be the number of active nodes, which we found to be generally much smaller than n . This makes the event detection scalable to large datasets. For instance the Wikipedia dataset contains 779005 nodes ($n = 779005$). However, the maximum number of active nodes in a pair of snapshots for the first 6 graphs is 300000, less than half of n . Table I gives the percentage of active nodes, for both datasets. It can be observed that the percentage of active nodes for a pair of snapshots never increases beyond 50% of the total number of nodes. We have incorporated the event-detection in a visual toolkit, where we used further optimizations to reduce the computation time, when the number of clusters is high [Yang et al. 2008]. The timing results for event detection on the DBLP and Wikipedia datasets are shown in Table I.

Fig 2 shows the variation in the number of merge and split events at different values of the κ parameter. As expected, the number of events reduce as κ is increased. It can be observed that after 0.5, the dip in the number of events is more pronounced. Thus one can adjust the κ parameter to a high value to capture only interesting merging or splitting clusters with a high degree of overlap.

We show the numbers of occurrences of the other events for the DBLP and Wikipedia datasets in Table II. From the table and Fig 2, we can observe that for DBLP, the *Form* and *Dissolve* events far outnumber the others. This indicates that most collaboration groups change quite drastically over time. In the case of Wikipedia, the *Continue* events are high. This is due to the fact that most pages and semantic links which are created are not changed much subsequently. The addition of new pages over time is captured by the increased *Form* events, while the relatively low number of *Dissolve* events are indicative of the reduced percentage of deletion in the Wikipedia webgraph.

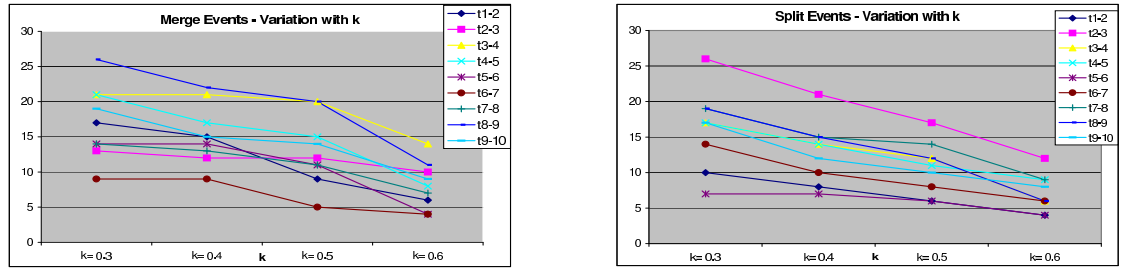


Fig. 2. Variation of merge and split events with parameter κ (DBLP dataset). The x-axis denotes different values of κ and the y-axis gives the number of occurrences.

Time	DBLP			Wikipedia		
	Continue	Form	Dissolve	Continue	Form	Dissolve
1-2	32	725	640	45	1106	2
2-3	33	677	680	267	2552	43
3-4	22	613	699	535	4534	272
4-5	27	792	600	699	3800	1499
5-6	19	555	806	2982	5785	3018
6-7	20	841	533			
7-8	24	828	773			
8-9	25	1058	787			
9-10	16	701	1054			

Table II. Event Occurrences for DBLP

5. BEHAVIORAL ANALYSIS

In this section, we show how the critical events can be employed to construct patterns for evolutionary behavior, and present some examples of useful behavioral measures. Most of the research on evolving networks [Backstrom et al. 2006; Falkowski et al. 2006] have focused solely on analyzing community behavior. In this section, we begin by presenting some movement-based measures to study the *behavior of nodes in the network* and their influence on others. In the second part, we describe how semantic information can be incorporated to aid the analysis of community-based events.

5.1 Movement-based Analysis

Movement for individuals is defined using the basic events - Join and Leave. We use these basic events to identify more complex behavior. In particular, we are interested in capturing the behavioral tendencies of individuals that contribute to the evolution of the graph. We wish to use these behavioral patterns to perform reasoning and predict future trends of the graph. We define four behavioral measures that can be *incrementally computed* at each time interval using the events discovered in the current interval. For each measure, we demonstrate its utility using a real-world application.

5.1.1 Stability Index. The Stability index measures the tendency of a node to have interactions with the same nodes over a period of time. A node is highly stable if it belongs to a very stable cluster, one that does not change much over time. Let $cl_i(x)$ represent the cluster that node x belongs to in the i^{th} time interval. The

Disease	Treat-ment	Age	Sex
diabetes +/- renal impairment	Drug	62	M
diabetes +/- renal impairment	Drug	59	M
hepatic impairment	Drug	56	M
diabetic neuropathy	Drug	66	F
diabetes +/- renal impairment	Drug	60	M
diabetes +/- renal impairment	Drug	62	F
diabetes +/- renal impairment	Drug	70	F
diabetes +/- renal impairment	Drug	66	M
diabetes +/- renal impairment	Drug	55	M
diabetes +/- renal impairment	Drug	50	M
diabetes +/- renal impairment	Drug	49	M
hepatic impairment	Drug	50	M
diabetic neuropathy	Drug	69	M
diabetes mellitus (type 2 niddm)	Drug	52	M
hepatic impairment	Drug	48	M
hepatic impairment	Drug	48	M
diabetes +/- renal impairment	Drug	49	M
hepatic impairment	Drug	49	M
diabetes mellitus (type 2 niddm)	Placebo	56	M

Table III. Low Stability Index - Clinical Trials Data

Stability Index (SI) for node x over T timestamps is measured incrementally as:

$$SI(x, T) = \sum_{i=1}^T \frac{|cl_i(x)|}{\sum_{j=1}^{V_i} (Leave(j, cl_i(x)) + Join(j, cl_i(x)))} \quad (4)$$

and $Activity(x) \geq Min_activity$.

Here $Activity(x) = \sum_{i=1}^T (x \in V_i)$ represents the number of time intervals x is active in. The threshold $Min_activity$ corresponds to the minimum number of active intervals for a node to be considered sociable. We used a $Min_activity$ value of $1/2$ the number of time intervals, for our experiments.

Stability for the Clinical Trials Dataset: In the case of the clinical trials data, nodes in a cluster correspond to individuals having similar observations. When a node has a low Stability index score, it indicates that the observations of that particular patient fluctuate appreciably. This causes the node to jump from one cluster to another repeatedly. This behavior represents an anomaly and can indicate possible side-effects of the drug being administered. Note that the dataset contains two groups of people, one group on the placebo and the other group on the drug with a distribution of 40:60. If the people with very low Stability index (outliers in this case) happen to be people on the drug, there is a reasonable indication that there may be a hepatotoxic effect from the drug intake, whereas if it is uniform over both sets of people, it would indicate there are no noticeable side-effects.

Accordingly, we computed the Stability index for all the nodes in the clinical trial data. On examination, we found 19 nodes having very low Stability index scores (below a threshold), indicating that these nodes move between clusters in almost every time interval that they are active. This suggests a significantly unstable behavior exhibited by these nodes. The 19 patients are shown in Table III.

In this particular application unstable nodes (patients) are a cause for concern since that may be an indication of toxicity. Drilling down on the nineteen most

unstable nodes we find that only one of them is on the placebo (rest were on the drug). In fact drilling down even further it is observed that out of the top 200 most unstable patients, eighty percent were on the drug. This indeed was very suspicious given the original distribution (40:60) and points to potential toxicity. *As it turns out, according to domain experts, this drug was discontinued for toxicity two years after this clinical trials study was conducted.*

5.1.2 Sociability Index. For the DBLP data, we define a related measure, called the Sociability Index. The Sociability Index is a measure of the number of different interactions that a node participates in. This behavior can be captured by the number of *Join* and *Leave* events that this node is involved in. Let $cl_i(x)$ be the cluster that node x belongs to at time i . Then, the Sociability Index is defined as:

$$SoI(x) = \frac{\sum_{i=1}^T (Join(x, cl_{i+1}(x)) + Leave(x, cl_i(x)))}{|Activity(x)|} \quad (5)$$

where $Activity(x) = \sum_{i=1}^T (x \in V_i)$ indicates the number of intervals that node x is active. Similar to the stability index, this is computed incrementally. The measure gives high scores to nodes that are involved in interactions with **different** groups.

Note that this measure does not represent the degree, which is a factor of the number of interactions a node is involved in. A case in point was a node that had a degree of 80, but a sociability of close to 0. When we examined the clusters the node belonged to, we found that the node interacted with the same nodes over several timestamps.

Application of Sociability for Cluster Link Prediction:

Problem Definition: The goal in link prediction [Liben-Nowell and Kleinberg 2003] is to use past interaction information to predict future links between nodes. Since, in this paper we are analyzing the evolution of clusters, our goal is to predict *future co-occurrences of nodes in clusters*.

In the case of the DBLP collaboration network, two nodes are clustered together if they work on related papers or belong to the same work-group, as we have seen. We use the behavioral patterns we have discovered to predict the likelihood of authors being clustered together in the future.

For prediction, we employ the Sociability index which, as we described before, gives the likelihood of an author being involved in different collaborations. If an author has a high Sociability Index score, the chances of him/her joining a new cluster in the future is very high. We compute the Sociability Index scores for authors using Equation 5. We use a degree threshold to prune these authors.⁷ We find all authors who have high Sociability index scores (> 0.75) and degree higher than the threshold and who have not been clustered together in the past. We then predict future cluster co-occurrences between them.

The seminal paper on link prediction [Liben-Nowell and Kleinberg 2003] provided an empirical analysis of several techniques for link prediction. We adopt the same

⁷The threshold value we used was 50 papers

scenario and split our DBLP snapshots into two parts. We use the clusterings for the first 5 years (1997-2001) to predict new cluster co-occurrences for the next 5 years. Note that we are only considering new links between authors. Hence we consider only authors that have not been clustered together previously. Also note that, the objectives here are not entirely the same as the above-mentioned paper, since we are looking to predict cluster co-occurrences rather than actual links between nodes.

Similar to the evaluation performed by Liben-Nowell and Kleinberg [2003], we use as our baseline a random predictor that randomly predicts pairs of authors who have not been clustered together before, and report the accuracy of all the methods relative to the random predictor. To perform comparisons, we implement three other approaches that were shown to perform well by the authors above:

Common Neighbor-based: This approach [Newman 2001; Liben-Nowell and Kleinberg 2003] gives high similarity scores to nodes that have a large number of neighbors in common. This measure is based on the notion that if two authors have a large number of common neighbors and have not yet collaborated, there is a good chance that they will, in the future. It is given by:

$$Score(a, b) = |\gamma(a) \cap \gamma(b)| \quad (6)$$

where $\gamma(a)$ represents the neighbors of node a .

Adamic-Adar: This measure, originally proposed by Lada et al. [2003] in relation to similarity between web pages, weights a common neighbor based on its importance. It is defined as:

$$Score(a, b) = \sum_{c \in (\gamma(a) \cap \gamma(b))} \frac{1}{\log(|\gamma(c)|)} \quad (7)$$

Nodes that have fewer neighbors are deemed more important than nodes with high degrees.

Jaccard coefficient: This measure, a popularly used similarity metric, computes the probability of two nodes having a common neighbor.

$$Score(a, b) = \frac{|\gamma(a) \cap \gamma(b)|}{|\gamma(a) \cup \gamma(b)|} \quad (8)$$

We used all the algorithms to predict cluster links for the last 5 years (2002-2006). We only considered pairs of authors who have not been clustered together in any of the 5 earlier snapshot graphs. The accuracy was computed as a factor of the random predictor [Liben-Nowell and Kleinberg 2003], which was found to give a correct result with probability 0.14%. The results are shown in Table IV. We find that the *Sociability Index-based method performs the best overall*, outperforming other approaches appreciably with a large ratio of correct predictions (275).

We believe that the Sociability index performs well in this application due to two reasons. First, it makes use of dynamic behavioral information, which the other measures are not designed to do. For the other neighborhood-based measures that

Predictor	Accuracy
Random Predictor Probability	0.14%
Sociability Index	275
Common Neighbors	25
Adamic-Adar	46
Jaccard Coefficient	23

Table IV. Cluster Link Prediction Accuracy. Accuracy score specifies the factor improvement over the random predictor. This method of evaluation is consistent with the one performed by Liben-Nowell and Kleinberg.

we considered, we had to create a cumulative graph with all the edges from 1996-2001 and use this static graph to find common neighbors. Since these measures do not take into account dynamic changes, their predictions are less efficient than the Sociability Index. Second, sociability is doing well for DBLP because collaborations are inherently influenced by sociability, which makes it a good heuristic for this particular dataset. Note that our goal in using this application is to show that our measure can capture the sociability behavior of nodes, and we are demonstrating this by showing its efficacy in predicting future cluster membership.

This result suggests that behavioral patterns of evolving graphs can be used to predict future behavior.

5.1.3 Popularity Index. The Popularity index is a measure defined for a cluster or community at a particular time interval. The Popularity Index of a cluster at time interval $[i, i + 1]$ is a measure of the number of nodes that are attracted to it during that interval. It is defined as:

$$PI(C_i^j) = \left(\sum_{x=1}^{V_i} Join(x, C_i^j) \right) - \left(\sum_{x=1}^{V_i} Leave(x, C_i^j) \right) \quad (9)$$

This measure is based on the transformation a cluster undergoes over the course of a time interval. If a cluster does not dissolve in $[i, i + 1]$ and a large number of nodes join the cluster and few leave it, then the cluster will have a high Popularity Index score. Note, that the Popularity index is an influence measure defined for a cluster. Also note, that this measure does not simply reflect the size of a cluster. However, the probability of a new node forming a link to at least one of the nodes in a cluster is proportional to the size of the cluster. Hence, larger clusters have higher propensity of attracting new nodes, which will cause more joins to these clusters, contributing to their popularity. This is illustrated in Figure 3 which shows the weak positive correlation between the size of clusters and their popularity scores. Note that, there are clusters which have high size and yet low popularity scores.

In the DBLP dataset, the popularity index can be used to find topics of interest for a particular year. For instance, if a large number of nodes join a cluster at a particular time point and a high percentage of them are working on a specific topic, it indicates a *buzz around that topic* for that year. On the other hand, if a large number of authors leave a cluster, and there are not many new nodes joining it, it indicates a loss of interest in a particular topic.

To find hot topics, we computed the popularity index scores for each cluster, and identified the most popular clusters, at each timestamp. We then examined the

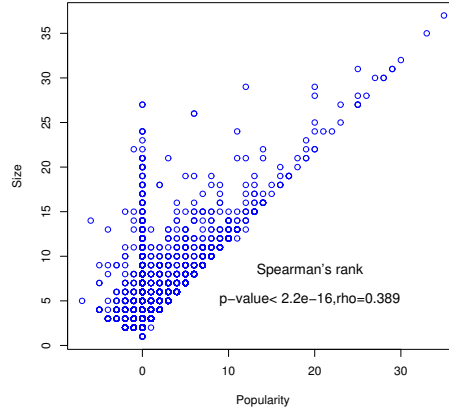


Fig. 3. Weak positive correlation between size and popularity scores on the DBLP datasets. Clusters over all timestamps are shown.

clusters that had high popularity scores to see if a large percentage of the authors in them were working on a particular topic.

We will now present an interesting result we obtained for the time span 1999-2000. In 1999, three authors Stefano Ceri, Piero Fraternali and Stefano Paraboschi formed a cluster. They were involved in a few papers on XML and web applications. In the next year (2000), these three authors were involved in a large number of collaborations, resulting in around 50 joins to their cluster. When we examined the topics of the papers that resulted, we found that 30 of these authors published papers related to XML. Since there were no papers on XML before 1999, this was a new and hot topic at that point. Since then there have been large number of papers on XML. Figure 4 shows the original 3 person cluster as well as the authors from the new cluster who were involved in XML related work in that particular time interval.

5.1.4 Influence Index. The influence index of a node is a measure of the influence this node has on others. Note that the influence that we are considering, in this case, is with regard to cluster evolution. We would like to find nodes that influence other nodes into participating in critical events. This behavior is measured for a node x , over all timestamps, by considering all other nodes that leave or join a cluster when x does. If a large number of nodes leave or join a cluster with high frequency when a certain node x does, it suggests that node x has a certain positive influence on the movement of the others. Let $Companions(x)$ represent all nodes over all timestamps that join or leave clusters with node x . The Influence for node x is given by:

$$Inf(x) = \frac{|Companions(x)|}{|Moves(x)|} \quad (10)$$

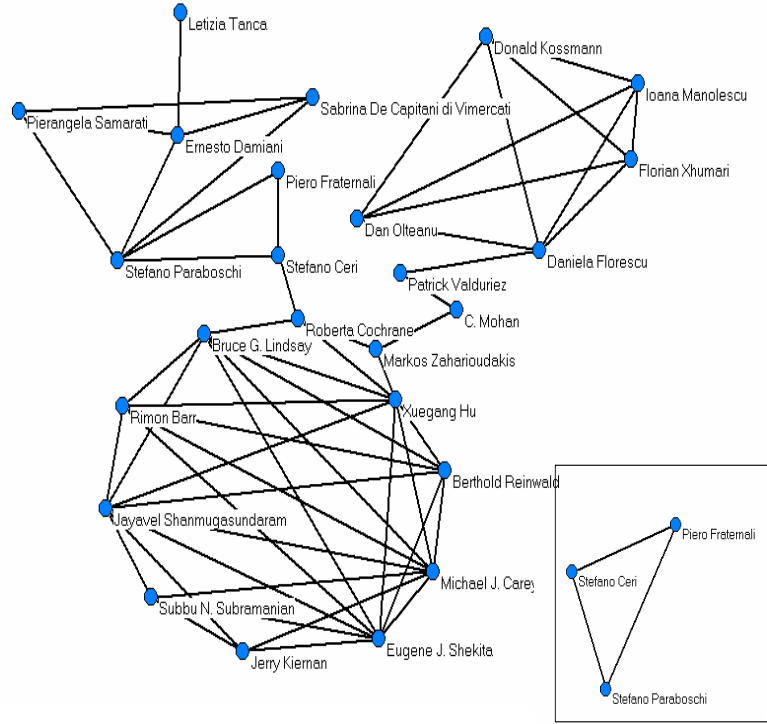


Fig. 4. Illustration of certain authors belonging to a very popular cluster (1999-2000 time period). Original cluster (3 authors) shown in small box. In the large graph, we show the connections among 25 authors from the new cluster who published XML-related papers in that time-frame.

Here $Moves(x)$ represents the number of *Join* and *Leave* events x participates in.⁸ Note that, this definition by itself, does not measure influence, since nodes that interact and move along with highly influential nodes will have high Influence score values as well. If we are interested in identifying only the most influential nodes, we need to eliminate these *follower* nodes. Hence, additional pruning constraints are needed.

Let $Max_Int(x)$ denote the node with which node x has the maximum number of interactions. Let $Deg(x)$ denote the number of neighbors of node x .

$Influence\ Index(x) = Inf(x)$ unless any of the following hold :

- $Inf(Max_Int(x)) > Inf(x)$
- $Deg(Max_Int(x)) > Deg(x)$

⁸To compute the Influence index efficiently, we incrementally update $Companions()$ and $Deg()$ for all nodes. The number of *Join* and *Leave* events ($Moves()$) are used in the Sociability case also and are stored incrementally as well.

Author	Influence Index
H. V. Jagadish	290.125
Hongjun Lu	268.5
Jiawei Han	266.625
Philip S. Yu	251.66
Rajeev Rastogi	246.85
Beng Chin Ooi	237
Tok Wang Ling	220.428
Heikki Mannila	206.5
Wenfei Fan	200.142
Qiang Yang	199
Johannes Gehrke	179.85
Christos Faloutsos	167.85
Rakesh Agrawal	157.875
Edward Y. Chang	153
Guy M. Lohman	131.375
Dennis Shasha	129.29
Jennifer Widom	128.375
Hamid Pirahesh	127.625
Michael J. Franklin	121.5
Hector Garcia-Molina	118.625

Table V. Top 20 Influence Index Values - DBLP Data

If any of the two conditions hold, $Influence\ Index(x) = 0$.

The additional constraints are imposed in order to ensure that we find the most influential nodes in the datasets. Note that, in some applications, when one is interested in identifying groups of influential nodes, the above pruning step may be eliminated. Accordingly, we have implemented the pruning step as a user-controllable flag in our program. We computed the Influence Index scores for nodes in the DBLP dataset. The top 20 authors are shown in Table V. We further illustrate the use of the Influence index in the next section.

In this subsection, we have presented different behavioral measures constructed from our basic events. Note that, it is possible for users to define other custom measures based on the events to capture and model other types of behavior efficiently.

5.2 Incorporating semantic content

For several real-world interaction networks such as online communities, WWW, collaboration and citation networks, an important factor influencing behavior and evolution is the semantic nature of the interactions itself. For instance, in a co-authorship network, two authors are connected if they publish a paper together. The topic or the subject area of the paper will definitely influence future collaborations for each of these authors. If there are different authors working on similar topics, the chances of them collaborating in the future is higher than two authors working on unrelated areas. In the case of Wikipedia, pages that are semantically related are linked together.

We wish to examine the influence of the semantics of the interaction on future interactions and incorporate semantic information for reasoning about evolution. Apart from reasoning, we also wish to develop measures for evaluating the events obtained from a semantic standpoint. For this purpose, we make use of both semantic category hierarchies and information theoretic measures.

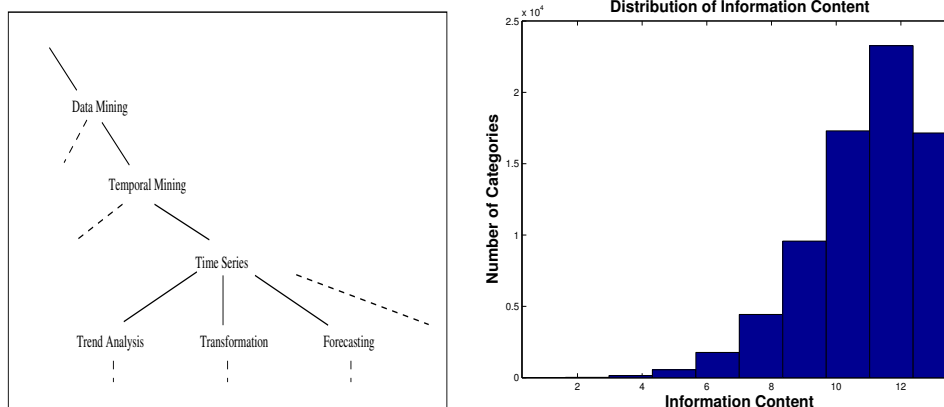


Fig. 5. a) A sample subgraph of the keyword DAG hierarchy. b) Distribution of Information Content for Wikipedia. The X-axis gives the IC values and the Y-axis represents the number of categories that have that particular value.

We use the DBLP co-authorship and the Wikipedia datasets for this analysis. For the DBLP dataset, to quantify the relevance of author pairs and construct a knowledge-base, we make use of semantic information from the topics of their papers. We begin by identifying a set of unique keywords composed of frequently used technical terms from the topics of all the papers in our corpus. We then group related keywords together to form keyword-sets, $k = \{w_1, \dots, w_n\}$, where each w_i is a keyword. Each paper can be labeled with a set of related keywords. An example of a keyword-set is $\{WWW, Web, Internet\}$. Thus each author can be associated with a paper-set, P , consisting of the union of the keyword-sets from all the papers that she/he co-authored in a particular time period. $P = \{k_1, \dots, k_p\}$ where each k_i is a keyword-set. However, the relationship between two authors cannot be inferred by merely comparing their paper-sets since different keywords are associated with different semantic meanings. One needs to consider the distribution of topics and the relationships among them.

To capture this, we constructed an ontology in the form of a hierarchy or a DAG where each node represents a keyword-set.⁹ Nodes at higher levels in the hierarchy represent keyword-sets that are more general, while nodes closer to the leaves represent more specific keywords. A node a has a child b if b is a keyword-set that represents a more specific term related to a . An example hierarchy is shown in Figure 5(a).

The Wikipedia dataset that we use [Gabrilovich and Markovitch 2007], already contains a category hierarchy comprising of 78664 categories. Most Wikipedia pages have more than one category associated with them. The categories of Wikipedia serve the same purpose as the keyword-sets detailed above. They provide information for classifying webpages (nodes) of the network according to their semantic relevance. We will refer to both of them (categories and keyword-sets) as terms,

⁹The keyword ontology was manually constructed by the authors for this work.

for the rest of this section.

Using these hierarchies, we can define the notion of semantic similarity [Lin 1998; Ganesan et al. 2003] in this context. To begin with, the Information Content (IC) of a term (category or keyword-set), using Resnik’s definition [Resnik 1999], is given as:

$$IC(k_i) = -\ln \frac{F(k_i)}{F(root)} \quad (11)$$

where k_i represents a term and $F(k_i)$ is the frequency of encountering that particular term over all the entire corpus. Here, $F(root)$ is the frequency of the root term of the hierarchy. Note that frequency count of a term includes the frequency counts of all subsumed terms in an is-a hierarchy. Accordingly, the root of our hierarchy includes the frequency counts of every other term in the ontology, and is associated with the lowest IC value. Note that terms with smaller frequency counts will therefore have higher information content values (i.e. more informative). The distribution of IC for the Wikipedia dataset is shown in Figure 5(b). It can be observed that most categories have high IC (low frequency) with only a few having low IC (high frequency).

Using the above definition, the Semantic Similarity (SS) between two terms (categories or keyword-sets) can be computed as follows:

$$SS(k_i, k_j) = IC(lcs(k_i, k_j)) \quad (12)$$

where $lcs(k_i, k_j)$ refers to the lowest common subsumer of terms k_i and k_j .

Next, we demonstrate how semantic similarity can be used to reason about community-based critical events, using examples from both datasets. To aid our analysis, we extend information theoretic measures such as information entropy and mutual information in this context.

5.2.1 Group Merge: The key intuition that we employ here is that the probability of a merge event depends on the Semantic Similarity between two clusters. For instance, if two clusters are comprised of authors working on highly related topics, it stands to reason that there is a high likelihood of a merge between them. We would like to investigate the strength of this relationship.

Let us consider two clusters C_i^a and C_i^b associated with k^a and k^b terms respectively. A simple way to compute the semantic similarity between two clusters is based on the information content of their terms, as shown below:

$$Inter_SS(C_i^a, C_i^b) = \frac{\sum_{k^a}^{m=1} \sum_{k^b}^{n=1} SS(m, n)}{k^a * k^b} \quad (13)$$

Clusters with high values of $Inter_SS()$, can be expected to contain authors or web-pages with similar topics and can hence be considered candidates for a possible merge in the future.

Semantic Mutual Information (SMI): To define the semantic similarity between two clusters, one can also employ an information theoretic mutual information measure [Strehl and Ghosh 2002]. Mutual information is a measure of the amount of statistical information shared between two distributions. We can compute the mutual information between two clusters based on the terms of all nodes

Cluster 1	Cluster 2	Merged Cluster
1) Querying Aggregate Data	Exact and Approximate Aggregation in Constraint Query	1) On the Content of Materialized Aggregate Views.
2) On the Orthographic Dimension of Constraint Databases		2) Automatic Aggregation Using Explicit Metadata
3) A Performance Evaluation of Spatial Join Processing Strategies		3) Reachability and Connectivity Queries in Constraint Databases

Table VI. Group Merge Event - Column 1 and 2 show the papers from the original clusters. Column 3 shows the papers from the merged cluster.

belonging to those clusters. This would be applicable when there is history data that can be used to estimate the probabilities. Given probabilities of terms m and n occurring in a cluster as $p(m)$ and $p(n)$ respectively, and their co-occurrence probability $p(mn)$, we can define the Semantic Mutual Information (SMI) between the two clusters C_i^a and C_j^b as:

$$SMI(C_i^a, C_j^b) = \sum_{m=1}^{k^a} \sum_{n=1}^{k^b} SS(m, n) * p(mn) * \log_{k^a * k^b} \frac{p(mn)}{p(m) * p(n)} \quad (14)$$

Note that each term pair is weighted by the information content of their most informative common ancestor in the keyword ontology (i.e. Lowest Common Subsumer).

In our implementation, we used the immediate history data, i.e the previous snapshot’s data, to compute the probabilities of terms (categories) as well as their co-occurrence probabilities ¹⁰.

The term information can be used to analyze the effect of topics on community based events. If two clusters in snapshot S_i - C_i^a and C_i^b , are merging in time T_i into cluster C_{i+1}^c , the inter-cluster similarity of C_i^a and C_i^b and their similarity to the new cluster formed, C_{i+1}^c can be indicative of the evolution of a topic by combining two similar sub-topics.

We illustrate this with an example from the DBLP dataset. In 1999, we found two clusters with high semantic similarity scores (shown in the first two columns of Table VI), due to the common keywords - ‘constraint’, ‘query’, ‘aggregate/aggregation’. We found that these two clusters merged in the following year (2000) giving a single cluster. All three clusters are shown in Table VI.

In the Wikipedia dataset, we found the merge events with highest SMI from each time stamp. We provide a couple of examples below from timestamps 2 and 3. Each of them depicts the two clusters that merge at those timestamps. The two merges each result in a single cluster with the main motifs being *Cinema* and *Chemical Elements* respectively ¹¹.

Cluster 1 at Time 2: Cinema of the United Kingdom ; Blowup ; Wilde ; Boris Karloff ; British Academy of Film and Television Arts ; Rowan Atkinson ; Time Bandits ; Ben Kingsley ; The Madness of King George ; Roger Moore ; School for Scoundrels ; Withnail and I ; My Beautiful Laundrette ;

¹⁰Note that it is possible to consider more history information to compute these probabilities

¹¹The merged cluster is large in both cases and hence not shown

Michael Caine ; Stan Laurel ; No Highway ; Whisky Galore! ; James Whale ; If... ; Born Free ; A Matter of Life and Death ; Albert Finney ; Neil Jordan ; John Boorman ; Richard Attenborough ; Our Man in Havana ; First Men in the Moon ; Brighton Rock ; The Happiest Days of Your Life ; Robert Morley ; Julie Walters ; The Man in the White Suit ; Truly Madly Deeply ; Mrs. Brown ; Billy Elliot ; David Lean ; Lionel Jeffries ; Forty-Ninth Parallel ; Topsy-Turvy ; Michael Powell ; My Left Foot ; The African Queen ; Irene Handl ; The Ipcress File ; Leslie Phillips ; Sid James ; Hayley Mills ; The Railway Children ; Joyce Grenfell ; Brassed Off ; Brief Encounter ; The Pope Must Die ; The Rebel ; The Lavender Hill Mob ; Chaplin ; James Mason ; A Canterbury Tale ; Michael Latham Powell ; Helen Mirren ; Walkabout ; Peter Sellers ; The Trials of Oscar Wilde ; Hugh Grant ; Straw Dogs ; Scrooge ; Blackmail ; John Schlesinger ; Ann Todd ; Alfie ; Billy Liar ; Danny Boyle ; Great Expectations ; The Ladykillers ; Little Voice ; Peter Greenaway ; Otley ; Genevieve ; The Man Who Fell to Earth ; Blithe Spirit ; The Day of the Jackal ; How I Won the War ; Bill Nighy ; British Independent Film Awards ; Jude Law ; Staggered ; Alec Guinness ; The Private Life of Sherlock Holmes ; The Winslow Boy ; Hattie Jacques ; Joan Sims ;

Cluster 2 at Time 2: George Lucas ; Plant ; Fish ; Cooking ; Food ; Osteichthyes ; Shark ; Species ; Nature (journal) ; Triumph of the Will ; International Energy Agency ; Fiji ; Pie ; Dicotyledon ; Economic and monetary union ; Black Narcissus ; Deathwatch ; The Entertainer ; My Learned Friend ; Vivien Leigh ; The First of the Few ; Juliet Mills ; The Curse of Frankenstein ; Dead of Night ; David Niven ; Mike Nichols ; The Goose Steps Out ; Elizabeth Hurley ; Oliver Twist ; A Room with a View ; The Third Man ; The Wooden Horse ; A Passage to India ; Phyllis Calvert ; The Wrong Arm of the Law ; Carry On Sergeant ; To Sir, with Love ; Doctor in the House ; Ned Kelly ; Jim Sheridan ; John Mills ; Dudley Moore ; Lindsay Anderson ; The Lady Vanishes ; The Adventures of Baron Munchausen ; Privilege ; The Full Monty ; 10 Rillington Place ; 84 Charing Cross Road ; Michael York ; Jenny Agutter ; Joan Collins ; Whistle Down the Wind ; The Italian Job ; The Wicker Man ; The Man Who Never Was ; Superman ; Deborah Kerr ; The Mummy ; A Fish Called Wanda ; An American Werewolf in London ; Dog Soldiers ; The Collector ; Herb ; Winter ; Constitutional monarchy ; Sturmabteilung ; Reduction ; New Guinea ; Spice ; Romanticism ; Altruism ; Jesus ; Kuru ; Olive ; Rose ; Hamlet ; Dill ; Onion ; Districts of Luxembourg ; Geography of Luxembourg ; Jean-Claude Juncker ; Luxembourg (district) ; Luxembourg (city) ; Lemon balm ; Mint ; Carolus Linnaeus ; Essential oil ; Candy ; Lamiales ; Annual plant ; Oregano ; Basil ; Chimpanzee ; Sandwich ; French cuisine ; Catalan language ; Classical music ; Sect ; Cannibalism ; Bay leaf ; Lao language ; Electronic media ;

Cluster 1 at Time 3: Chemist ; Gadolinium ; Gadolinite ; Fluoride ; Rare earth ; Period 6 element ; Geologist ; Lanthanide ; Sulfide ; Thermal neutron ; Hexagon ; Bromide ; Selenide ; Terbium ; Johan Gadolin ; Metabolism ; Didymium ; Compact disc ; Chloride ; Alpha decay ; Xenon ; Magnetic resonance imaging ; Dysprosium ; F-block ; Bastnasite ; Monazite ; Nuclear control rod ; Toxicity ; Nitride ; Iodide ; Infrared ; Holmium ; Marc Delafontaine ; Garnet ; Laser ; Erbium ; Absorption band ; Carl Gustaf Mosander ; Magnetic moment ; Ytterbium ; Thulium ; Niobe ; Euxenite ; Period 7 element ; Vaxholm ; Greek ; Nuclear energy ; Xenotime ; Fergusonite ; Polycrase ; Chalcogenide ; Nuclear fusion ; High Flux Isotope Reactor ; Oak Ridge National Laboratory ; Los Alamos National Laboratory ; 1 E4 s ; Anhydrous ; Filter ; Charles James ; Dopant ; Ultraviolet ;

Cluster 2 at Time 3: Antoine Lavoisier ; Year ; Acid ; Oxygen ; Fluorine ; Bullet ; Atomic number ; Scientific notation ; Electron ; Boiling point ; Half-life ; Chemical series ; Specific heat capacity ; Ohm ; Decay product ; Kilogram per cubic metre ; Ionization potential ; Thermal conductivity ; Natural abundance ; Chemical element ; Crystal structure ; Glass ; Decay energy ; Periodic table ; Energy level ; Neutron ; Zinc ; Atomic radius ; Speed of sound ; Si ; Covalent radius ; Periodic table (standard) ; Melting point ; Stable isotope ; Vapor pressure ; Molar volume ; Decay mode ; Electronegativity ; Metre per second ; Electron configuration ; List of elements by name ; Isotope ; Kelvin ; Electrical conductivity ; Periodic table block ; Periodic table period ; Periodic table group ; Atomic mass unit ; Kilojoule per mole ; Mega ; List of elements by symbol ; Van der Waals radius ; Metal ; Color ; Pascal ; Uranium ; Nuclear weapon ; Radioactive ; Calcium ; 1 E-25 kg ; Europium ; Germanium ; Dmitri Mendeleev ; P-block ; Distillation ; Transistor ; Gram ; Silicon ; Gallium ; Crystal ; Iodine ; Einsteinium ; Indium ; Rectifier ; Thermistor ; Indigo ; Ferdinand Reich ; True metal ; Solder ; Thallium ; Poor metal ; Welding ; Deuterium ; Tungsten ; Light bulb ; Cubic metre ; Spallation ; 1811 ; Sodium ; Alkali metal ; Lithium ; Lutetium ; Cerium ; Boron ; Technetium ; Niobium ; Potassium permanganate ; Steelmaking ; Manganese nodule ; Arginase ; Dry cell ; Cobalt ; Neodymium ; Enamel ; Promethium ; Rocket ; Tantalus ; Pyrochlore ; 1 E8 s ; 1 E10 s ; 1 E15 s ; 1 E11 s ; Silicate ; Neutron emission ; Heinrich Rose ; Isomeric transition ; Superconducting magnet ; Capacitor ; Charles Hatchett ; Osmium ; Phonograph ; Fountain pen ; Shocked quartz ; Dinosaur ; Fingerprint ; Hassium ; Artificial pacemaker ; Smithson Tennant ; Neon ; Granite ; Neptunium ; Philip Abelson ; Transmutation ; Edwin McMillan ; Protactinium ; Plutonium ; University of California ; X-ray ; Gamma ray ; Scandium ; Electrode ; Sulphur (disambiguation) ; Czochralski process ; Feldspar ; Humphry Davy ; Opal ; Jasper ; Zone melting ; Amethyst ; Diatom ; Period 3 element ; Meteoroid ; Mass number ; Silane ; Trichlorosilane ; Hornblende ; Zinc chloride ; Silicon tetrachloride ; Semiconductor device ; Tektite ; Chemical equation ; Abrasive ; Solar cell ; Turbine ; reaction ; Sodium cyanide ; Bleach ; Chlorite ; Mustard gas ; Water purification ; Synthetic rubber ; Critical temperature ; Critical exponent ; Isotherm ; Miscibility ; Check

valve ; John Ambrose Fleming ; Barium oxide ;

The above examples illustrate the relationship between semantic similarity between clusters and the *Merge* event. In both cases, we can observe that the merging clusters are highly similar in their content. The SMI measure defined above can thus be used to evaluate the semantic relevance of merges. The Wikipedia encyclopedia evolves by linking pages with similar content as well as constructing new pages. Hence, the Merge events that we are capturing do convey information regarding the community structure and evolution of the webpages.

We have seen how high SMI merges provide semantic justification for our event-detection process. On the other hand, SMI also allows one to identify unexpected merges, which can be quite interesting. These are merges with lower SMI, and hence represent the convergence of clusters with lower semantic similarity.

For instance, let us consider a cluster merge event for the DBLP dataset, that occurred in the 2005-2006 time interval. Our algorithm identified two groups (one from Germany and one from Italy) who independently published articles in different conferences in 2005.

Cluster 1 at Time 9:

AAAI 2005 : *Niels Landwehr, Kristian Kersting, Luc De Raedt*: nFOIL: Integrating Nave Bayes and FOIL

AAAI 2005 : *Luc De Raedt, Kristian Kersting, Sunna Torge*: Towards Learning Stochastic Logic Programs from Proof-Banks.

Cluster 2 at Time 9:

ICML 2005 : *Sauro Menchetti, Fabrizio Costa, Paolo Frasconi*: Weighted Decomposition Kernels.

IJCAI 2005 : *Andrea Passerini and Paolo Frasconi*: P. Kernels on Prolog Ground Terms.

Merged Cluster at Time 10:

ILP 2006 : *Niels Landwehr, Andrea Passerini, Luc De Raedt, Paolo Frasconi*: kFOIL: Learning Simple Relational Kernels

The original clusters had low SMI, due to the fact that the corresponding authors were working on reasonably diverse topics, with Niels Landwehr and Luc De Raedt working on Inductive Logic and Passerini and Frasconi, working on kernels. However, they still collaborated together to combine their ideas¹². The authors describe their joint work as ‘*A novel and simple combination of inductive logic programming with kernel methods is presented. The kFOIL algorithm integrates the well-known inductive logic programming system FOIL with kernel methods.*’

Also, when we further analyzed the merge events which had lower SMI values, we found that although some of the corresponding clusters contained nodes associated with suitably disparate categories, they were being clustered together due to their *being linked by a common neighbor*. This caused the clusters to merge, although their semantic similarity was fairly low. These kind of merges can also be interesting, since they can help one to uncover hidden relationships across nodes. This is related to the notion of Sociability that we examined in the previous subsection. Sociable nodes are nodes that interact with very different nodes. This can sometimes lead to the disparate nodes themselves getting involved together. Note that

¹²One possible reason for the merge could be their locations.

Cluster	Split Cluster 1	Split Cluster 2
1) Web Site Evaluation: Methodology and Case Study	Spatio-temporal Information Systems in a Statistical Context.	RoadRunner: automatic data extraction from data-intensive web sites.
2) RoadRunner: Towards Automatic Data Extraction from Large Web Sites.		
3) SIT-IN: a Real-Life Spatio-Temporal Information System.		

Table VII. Group Split Event - Column 1 shows the papers from the original cluster. Columns 2 and 3 show the papers from the split clusters.

we observed this behavior more with DBLP than Wikipedia. For the Wikipedia dataset, all merges had reasonably high SMI values. This is due to the fact that in Wikipedia, pages are linked due to similar semantic content. Hence, merges are likely to have a high semantic context. For DBLP, our analysis is limited by the hierarchy that we have manually constructed, using only paper titles. It is entirely possible that all semantic relationships are not captured.

One relatively straightforward conclusion we could make from our observations in this and the previous subsection is that the propensity of a merger between clusters is dependent on two main factors - the *sociability of the nodes* and the *semantic similarity of the terms* involved. We have presented measures to capture both these types of behavior.

5.2.2 *Group Split*:. We believe that an important factor for a *Split* event is topic divergence. The probability of topic divergence is inversely proportional to the semantic similarity between topics in a cluster. We can define the intra-cluster semantic similarity *Intra_SS* as the average of the semantic similarity scores between the keyword-sets in the cluster. The semantic similarity within a cluster is given by:

$$Intra_SS(C_i^a) = \frac{\sum_{k^a}^{m=1} \sum_{k^a}^{n=m+1} SS(m, n)}{k^a * (k^a - 1)} \quad (15)$$

where k^a , as before, represents the total number of terms in cluster C_i^a . If the intra-cluster semantic similarity is small, then it indicates that the cluster is likely to split in the next few timestamps. For instance, in 2001 we found a cluster with relatively low intra-cluster semantic similarity. This cluster contained very disparate keyword-sets $\{\{\text{web,data extraction}\}\}$ and $\{\{\text{spatio-temporal, information system}\}\}$. The papers in this cluster are shown in the first column of Table VII. In 2002, this cluster split into two different clusters, shown in the columns 2 and 3 in Table VII. Thus, the semantic similarity within clusters can be indicative of possible future *Split* events.

Semantic Information Gain (SIG): We can also define an information theoretic measure to analyze a Split event, using the notion of Information Gain. Information Gain is popular in use in decision trees (C4.5), to identify splitting attributes. It measures the difference in entropy between the original distribution and the current distribution. In the case of a Split event, it can be used to measure the change in entropy in terms of categories of the original cluster and the two split clusters. The key intuition here, is that if a Split occurs due to the divergence of topics

or categories, the entropy of the resultant clusters will be much lower than the original cluster. Assume a cluster C_i^a splits into two clusters at time $i + 1$, C_{i+1}^b and C_{i+1}^c . Let K_a be the number of terms in the original cluster, and let K_b and K_c be the number of these terms that are carried through to the two split clusters. The entropy for a cluster C_i^a based on the terms K_a it is associated with, can be given as:

$$H(C_i^a) = - \sum_{m=1}^{K_a} IC(m) * p(m) * \log\left(\frac{1}{p(m)}\right) \quad (16)$$

where $p(m)$ represents the probability of term m occurring in a cluster. Note that we are weighting the entropy by the Information Content of the terms involved. If a cluster is associated with multiple different categories, it can have high entropy within itself. The Semantic Information Gain for a Split event can be computed based on the entropies of the three clusters as :

$$SIG(C_i^a, C_{i+1}^b, C_{i+1}^c) = H(C_i^a) - \left(\frac{K_b}{K_a}H(C_{i+1}^b) + \frac{K_c}{K_a}H(C_{i+1}^c)\right) \quad (17)$$

A high value of Semantic Information Gain is indicative of divergence of topics, as the two resultant clusters will be more aligned to certain terms than others. By studying these Split events, one can discern important branches in the evolution of topics over time. Note that, although we have presented the notation above for a 2-way Split event, it can be generalized to multi-way splits. Also note that, in the above formulation, we are evaluating the Split based only on the terms of the original cluster. Hence we are not considering the new nodes that might belong to these new clusters and their associated terms.

We illustrate the efficacy of this measure with a few examples from the Wikipedia dataset. For each Split event that we discovered, we computed the resulting Semantic Information Gain. We analyze the ones with the highest and lowest SIG values.

High Semantic Information Gain : Below, we show a 3-way Split event with high SIG obtained between snapshots 4 and 5 of the Wikipedia dataset. The original cluster at time 4 contained pages on 3 different topics - folk music, Philippines and Mississippi. The resulting 3 clusters at time 5 comprise of pages belonging to each of these categories.

Cluster at Time 4: Folk music, Zouk, Milonga ; Lambada ; Tsifteteli ; Christine Lavin ; Hula ; Meringue ; Plena ; Kunqu ; Cante jondo ; Qawwali ; Lam ; Country dancing ; Guajira ; Bachata ; Taarab ; Reel ; Maqam ; Batucada ; Yodeling ; The Dubliners ; Shango ; Parang ; West gallery music ; Chimurenga ; Kalinda ; Muqam ; The Weavers ; Mbira ; Fandango ; Sawt ; Choro ; The Chieftains ; Tambu ; Francis James Child ; Bangsawan ; Folk clubs ; Gwo ka ; Candombe ; Pentangle ; Beguine ; Forro ; Tango ; Dangdut ; Giddha ; Planxty ; Oro ; Gagaku ; Honkyoku ; Seamus Ennis ; Rada ; Kheyal ; Farruca ; Dastgah ; Donegal fiddle tradition ; Clannad ; Tarantella ; Waitata ; Carimbo ; Fairport Convention ; Compas ; Thumri ; Jig ; Clare ; Halling ; Shabad ; Tsamiko ; Sid Kipper ; Zydeco ; Fado ; Klezmer ; Timba ; Mississippi ; Hattiesburg, Mississippi ; Vicksburg, Mississippi ; Alternative political spellings ; Pontotoc, Mississippi ; Ackerman, Mississippi ; McComb, Mississippi ; Starkville, Mississippi ; Columbus, Mississippi ; Laurel, Mississippi ; Iuka, Mississippi ; Gulfport, Mississippi ; Blue Mountain College ; Jackson State University ; University of Mississippi Medical Center ; Mississippi College ; Corinth, Mississippi ; Natchez, Mississippi ; Gulf Islands National Seashore ; Belhaven College ; Hurricane Camille ; Tombigbee River ; Mississippi University for Women ; Magnolia Bible College ; Tougaloo College ; Yazoo River ; Adams County, Mississippi ; Alcorn State University ; Delta State University ;

Brookhaven, Mississippi ; Cat Island ; Kosciusko, Mississippi ; Tishomingo County, Mississippi ; Wayne County, Mississippi ; Yazoo County, Mississippi ; University of Southern Mississippi ; Mississippi Valley State University ; Horn Island ; Natchez Trace Parkway ; William Carey College ; Foreign relations of the Philippines ; Transportation in the Philippines ; Communications in the Philippines ; Economy of the Philippines ; Ilocos Region ; EDSA Revolution ; Flag of the Philippines ; Central Visayas ; Eastern Visayas ; Western Visayas ; Central Luzon ; Bicol Region ; Cagayan Valley

Cluster 1 at Time 5: Muezzin ; Philippine peso ; Foreign relations of Peru ; Military of Peru ; Metro Manila ; Lupang Hinirang ; Philippine Sea ; Autonomous Region in Muslim Mindanao ; Geography of the Philippines ; Davao Region ; Caraga ; Zamboanga Peninsula ; Northern Mindanao ; SOCCSKSARGEN ; Gloria Macapagal-Arroyo ; Filipino ; Transportation in the Philippines ; Ilocos Region ; Central Visayas ; Flag of the Philippines ; Bicol Region ; Central Luzon ; Cagayan Valley ; Cordillera Administrative Region ; Eastern Visayas ; Western Visayas

Cluster 2 at Time 5: Choctaw ; Mississippi ; Flag of Mississippi ; Gulfport, Mississippi ; Hattiesburg, Mississippi ; Greenville, Mississippi ; Millsaps College ; Pontotoc, Mississippi ; Ackerman, Mississippi ; McComb, Mississippi ; Starkville, Mississippi ; Tupelo, Mississippi ; Laurel, Mississippi ; Iuka, Mississippi ; Blue Mountain College ; University of Mississippi Medical Center ; Mississippi College ; Corinth, Mississippi ; Natchez, Mississippi ; Gulf Islands National Seashore ; Hurricane Camille ; Mississippi State University ; Tombigbee River ; Woodall Mountain ; Mississippi University for Women ; Magnolia Bible College ; Yazoo River ; Adams County, Mississippi ; Alcorn State University ; Delta State University ; Brookhaven, Mississippi ; Cat Island ; Pascagoula, Mississippi ; Kosciusko, Mississippi ; Tishomingo County, Mississippi ; Wayne County, Mississippi ; Yazoo County, Mississippi ; University of Southern Mississippi ; University of Mississippi ; Mississippi Valley State University ; Horn Island ; Natchez Trace Parkway ; William Carey College ; Choctaw mythology

Cluster 3 at Time 5: Folk music ; Dhrupad ; Milonga ; Lambada ; Tsifteteli ; Hula ; Meringue ; Plena ; Kunqu ; Cante jondo ; Qawwali ; Lam ; Country dancing ; Bachata ; Taarab ; Reel ; Batucada ; Yodeling ; The Dubliners ; Shango ; Parang ; West gallery music ; Chimurenga ; Kalinda ; Muqam ; The Weavers ; Berimbau ; Merengue ; Mbira ; Bhajan ; Sawt ; Choro ; Tambu ; Bangsawan ; Folk clubs ; Gwo ka ; Candombe ; Beguine ; Forro ; Dangdut ; Giddha ; Planxty ; Oro ; Honkyoku ; Seamus Ennis ; Rada ; Tarana ; Kheyal ; Sea shanty ; Dastgah ; Clannad ; Tarantella ; Waiata ; Carimbo ; Doina ; Fairport Convention ; Compas ; Thumri ; Halling ; Shabad ; Tsamiko ; Sid Kipper ; Zydeco ; Fado ; Klezmer ; Timba ; Blood on the Tracks ; Macarena

It can be observed that the new clusters are more semantically related than the original cluster. Furthermore, they contain other nodes which were not in the original cluster, that relate to the central motifs of the cluster. This again illustrates the evolution of the Wikipedia encyclopedia. As new pages are created, the overall structure is improved, with semantically grouped communities of webpages.

Low Information Gain : As is evident from the above example, high SIG values can indicate semantically meaningful splits. On the other hand, low SIG values can indicate subtle changes across snapshots, where a cluster splits into two parts due to a small semantic difference among the associated categories. These splits can be interesting as they can reveal differences that may not be obvious. This can be considered akin to drilling down a hierarchy to discover subtle specializations of a category. We demonstrate this type of evolutionary behavior below with a couple of examples from the Wikipedia dataset.

Cluster at Time 1: Algebra ; Linear equation ; Quadratic equation ; Scalar ; System of linear equations ; Superposition ; Linear function ; Cubic equation ; Function (mathematics) ; Quartic equation ; Quintic equation ; Line (mathematics) ; Identity

Cluster 1 at Time 2: Quadratic equation ; Fundamental theorem of algebra ; Loss of significance ; Complex conjugate ; Cubic equation ; Quartic equation ; Root (mathematics) ; Brahmagupta ; Quintic equation ; Completing the square ; Quadratic irrational ; Abraham bar Hiyya Ha-Nasi ; Al-Khwarizmi

Cluster 2 at Time 2: Algebra ; Linear equation ; Scalar ; System of linear equations ; Superposition ; Linear function ; Function (mathematics) ; Line (mathematics)

We can observe that the cluster on algebra and equations at Time 1, has split into two clusters specializing in information regarding Quadratic and Linear equations respectively. This change has low SIG, since both the new clusters are strongly related to each other and the original cluster. A more ironic example, given below, shows two new clusters on East and West Berlin.

Cluster at Time 2: August 13 ; Berlin Wall ; East Berlin ; November 9 ; October 3 ; West Berlin ; German reunification ; Boroughs of Berlin ; Pankow ; Lichtenberg ; Prenzlauer Berg ; Friedrichshain ; Treptow

Cluster 1 at Time 3: November 9 ; October 3 ; Sovereignty ; West Berlin ; German reunification ; History of Germany since 1945 ; Spandau ; Kreuzberg ; Boroughs of Berlin ; Charlottenburg ; Wilmersdorf ; Tiergarten ; Berlin S-Bahn ; Judgment in Berlin ; Stunde Null ; Zehlendorf, Berlin

Cluster 2 at Time 3: East Berlin ; Pankow ; Lichtenberg ; Prenzlauer Berg ; Friedrichshain ; Treptow

In both the above cases, the original cluster was not semantically diverse. It contained pages that were similar. However, the newly formed clusters reflect a further improvement in the overall entropy, and the evolution of the encyclopedia into more meaningfully linked communities of webpages. This kind of behavior can be captured using the SIG.

For the DBLP dataset, we found a Split event with low SIG, involving a cluster consisting of papers on structure extraction from HTML and unstructured documents.

Cluster at Time 3

FODO 1998: *Seung Jin Lim, Yiu-Kai Ng: Constructing Hierarchical Information Structures of Sub-Page Level HTML Documents*

ER 1998: *David W. Embley, Douglas M. Campbell, Y. S. Jiang, Stephen W. Liddle, Yiu-Kai Ng, Dallan Quass, Randy D. Smith: A Conceptual-Modeling Approach to Extracting Data from the Web.*

IDEAS 1998: *Aparna Seetharaman, Yiu-Kai Ng: A Model-Forest Based Horizontal Fragmentation Approach for Disjunctive Deductive Databases*

CIKM 1998: *David W. Embley, Douglas M. Campbell, Randy D. Smith, Stephen W. Liddle: Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents*

In the next year (1999), this cluster splits into two different clusters. While Seung Jin Lim, Yiu-Kai Ng and David W. Embley continue working on extracting information from Web Documents, Stephen W. Liddle, Douglas M. Campbell, Chad Crawford specialized on Business Reports.

Cluster 1 at Time 4

CIKM 1999: *Seung Jin Lim, Yiu-Kai Ng: An Automated Approach for Retrieving Hierarchical Data from HTML Tables.*

DASFAA 1999: *Seung Jin Lim, Yiu-Kai Ng: WebView: A Tool for Retrieving Internal Structures and Extracting Information from HTML Documents*

SIGMOD 1999: *David W. Embley, Y. S. Jiang, Yiu-Kai Ng: Record-Boundary Discovery in Web Documents.*

Cluster 2 at Time 4

CIKM 1999: *Stephen W. Liddle, Douglas M. Campbell, Chad Crawford: Automatically Extracting Structure and Data from Business Reports.*

Our observations over the course of this subsection have revealed the benefits of the entropy-based measure to analyze Split events. First, it provides us a means to evaluate and justify the semantic relevance of *Split* events that the event detection

Cluster 1	Cluster 2
Object Recognition Using Appearance-Based Parts and Relations	Hierarchical Organization of Appearance-Based Parts and Relations for Object Recognition
Mining Insurance Data at Swiss Life	A Data Mining Support Environment and its Application on Insurance Data
M-tree: An Efficient Access Method for Similarity Search in Metric Spaces	Processing Complex Similarity Queries with Distance-Based Access Methods
Optimizing Queries in Distributed and Composable Mediators	Distributed View Expansion in Composable Mediators
Scaling up Dynamic Time Warping to Massive Dataset to Massive Datasets	Scaling up dynamic time warping for datamining applications

Table VIII. Continue Events - Column 1 shows a paper from a cluster that is part of a Continue event. Column 2 shows the paper from the cluster in the next timestamp. In each case we can observe the evolution of topics and ideas from the extensions to earlier papers.

algorithm discovers. Second, our analysis has shown how key information regarding community evolution, such as topic divergence, can be gleaned using this measure.

5.2.3 *Group Continue*:. For a Continue event, since the nodes belonging to the cluster do not change, one can ascertain information about how ideas and links evolve. This is of more relevance to the DBLP dataset, where one can study the evolution of topics as well as collaborations. The clusters that correspond to a continue event will tend to have a reasonably high SMI score. Note that, in this case, we are measuring SMI of the same cluster across successive snapshots, rather than two clusters in the same snapshot as in the Merge case. We present some examples of the papers corresponding to continue events for the DBLP dataset in Table VIII. The first column in the table represents a paper from the cluster at time stamp i and the second column denotes the most similar paper from the continuing cluster at $i + 1$. As we can observe, there is a marked progression in the topics of papers over time.

6. DIFFUSION MODEL FOR EVOLVING NETWORKS

We use the behavioral patterns discussed in the previous section to define a diffusion model for evolving networks. Diffusion models have been studied for complex networks [Alkemade and Castaldi 2005; Cowan and Jonard 2004] and specifically in the context of influence maximization [Kempe et al. 2003; 2005] where the task is to identify key start nodes that can be used to effectively propagate information through the network. The information can be either an idea or an innovation that propagates through the network over time. In this regard, Kempe et al. [2003; 2005] discuss two models for the spread of influence through social networks. We examine this scenario from an evolving perspective, where the nodes and edges of the network are transient.

Let us consider an idea or innovation that arrives into the network at timestamp a . We define four states for nodes in the evolving network - *active*, *inactive*, *contagious* and *isolated*. These states are not mutually exclusive, as we will see later. At the beginning of the diffusion process, at time a , all nodes in the network are *inactive*. The diffusion model begins with a set of nodes that are activated

(provided the information) at the first timestamp. These *active* nodes will be *contagious* briefly, in that, in the next timestamp they can activate other nodes they interact with, passing on the information they received. Subsequently, the newly *contagious* nodes proceed to attempt to activate their *inactive* neighbors. The process continues, with the information propagating through the network until at time T there are $\sigma(T)$ active nodes in the network. In earlier work, the effect of a *contagious* node has been limited to one timestamp, which means that an *active* node can attempt to activate its neighbors only once. However this does not capture the fact that the network topology can change, with the neighbors of nodes changing over time. After a *contagious* node has activated some of its neighbors, new nodes might come in contact with it in subsequent time instances. In this regard, we relax this constraint allowing a node to remain *contagious* when confronted with new neighbors. A node can thus attempt to activate each unique neighbor once.

When a node is surrounded by *contagious* nodes, its propensity to get activated is given by an activation function.

Definition: The activation function for a node v , $Ac_v()$ is a non-negative function that maps the weights associated with the neighbors of v , $wt(x, v) \forall x = neighbor(v)$ to either 0 or 1.

We describe two Activation functions, *Max* and *Sum*, for a node v as

$$Ac_v^{max}(u_1, u_2, \dots, u_m) = (\arg \max_{1 \leq i \leq m} (wt_v(u_i)) \geq \theta_v) \quad (18)$$

$$Ac_v^{sum}(u_1, u_2, \dots, u_m) = \sum_{1 \leq i \leq m} wt_v(u_i) \geq \theta_v \quad (19)$$

Here, θ_v denotes the activation threshold for node v . The weights on the edges represent the likelihood of that particular interaction leading to an activation. If the edge between two nodes has a high weight, it indicates that if one of the nodes gets activated, the chance of it activating the other is high. In our case, we define the weights for an interaction based on the Sociability Index values of the nodes involved, since Sociability can best capture the aforementioned property. If a node is highly sociable, it has a high propensity of passing on information to other nodes it interacts with. Hence, for each interaction of node x with a neighbor, y , the weight of the interaction is given by

$$wt_x(y) = SoI(y) \quad (20)$$

Similarly $wt_y(x) = SoI(x)$. Note that since we are dealing with diffusion over time, the $SoI(x)$ represents the cumulative value defined in (5) until the current time point. The Sociability values thus can change over time.

The set of nodes activated in a given time interval i due to the initial node x and the cardinality of this set are given by $R_x(i)$ and $\sigma_x(i)$ respectively. The total set and number of nodes activated due to x after T timestamps of the diffusion process are given as

$$R_x(T) = \cup_{i=1}^T R_x(i) \quad (21)$$

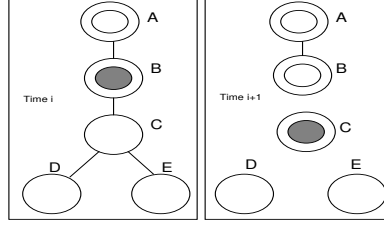


Fig. 6. Isolation of active nodes. The double circles indicate active nodes. The grey inner circle represents contagious nodes. Nodes D and E are inactive.

$$\sigma_x(T) = \sum_{i=1}^T \sigma_x(i) \tag{22}$$

It is also important to consider the effect of deleted nodes and edges. When a node is not participating in any interaction in the current timestamp it is said to be *isolated*. An *isolated* node cannot influence any other nodes since it has no interactions.

CLAIM 6.1. *An active node can be isolated.*

PROOF. As we mentioned earlier, the topology of the network can change at every timestamp. Hence, a node that has just become active can be separated from its neighbors due to the deletion of edges. The node will then remain isolated until a new interaction is formed with it. □

An example of this scenario is shown in Figure 5. Node *A* begins the diffusion process activating node *B*. *B* is contagious at time *i* and activates node *C*. However at the next timestamp, *C* no longer interacts with *B*, *D* and *E*. Although it is active and contagious, it is isolated at this time instant. In the future, if it interacts with other nodes, it can attempt to activate them once.

Influence Maximization: Influence Maximization is an important problem for diffusion models and has practical applications in viral marketing and epidemiology. The challenge is to find an initial set of active nodes that can influence the most number of inactive nodes over the duration of the diffusion.

Problem Definition: Given a graph *G* that evolves over *T* timestamps and a diffusion model, the task is to find the set of *k* initial nodes *S* to maximize $R_S(T)$ where $R_S(T) = \cup_{x \in S} R_x(T)$

Kempe et al. [2003; Kempe et al. [2005] discuss a greedy algorithm for finding the initial set that maximizes the influence. They find the start nodes that maximize $\sigma(T)$, where $\sigma(T) = \sum_{x \in S} \sigma_x(T)$. To find $\sigma_x(T)$ for all nodes *x*, they simulate the diffusion process over the network. However, in our case, the network is dynamic with edges and nodes getting added or deleted. At a particular timestamp *i*, it is unclear how the network is going to change at time *i* + 1. Hence, simulating the diffusion on the static graph will not work. Considering high-degree nodes to start the diffusion process has been examined in social network research [Wasserman and Faust. 1994]. However, using the degree to determine the initial nodes may not be a good option [Kempe et al. 2003], since it is possible for nodes of high degree to be

Method	Activated nodes (%)	
	Max Activation	Sum Activation
Random	16.67	20.39
Accumulated Degree	51.9	65.33
Influence	61.12	81

Table IX. Diffusion Results

clustered, which limits their range. Instead, we advocate the use of the Influence Index we defined in the previous section for this purpose. The Influence Index is an incremental measure which considers the behavior of the nodes over the previous timestamps and chooses nodes that have the highest degree of influence over other nodes. Also, *by pruning followers of influential nodes*, we are ensuring that the nodes with high influence index are *not likely to be clustered*.

Empirical Evaluation: We conducted an experiment to evaluate the performance of the Influence index-based initialization. To compare, we employed an approach based on accumulated degree, where we picked nodes that had the highest degree, over the preceding timestamps, to be the start nodes. As a baseline, we implemented a random approach where the initial nodes are chosen at random. We constructed a graph using a subset of nodes from the DBLP collaboration network. We considered the interactions from 1997-2001 to compute sociability, degree and influence scores. We then assumed the introduction of a new idea at 2002 and then tracked its diffusion through the network over the next 4 timestamps (till 2006). We used an active set size, k , of 5 and both the Sum and Max activation functions. We performed the experiments 100 times, choosing random activation thresholds for the nodes from $[0,1]$. The results are shown in Table IX. Our results suggest that the Influence index can be useful in this regard. It succeeds in *activating 61% and 81% of the nodes* in the network in 4 timestamps for the Max and Sum Activation functions respectively, clearly outperforming the other approaches.

7. DISCUSSION

In this paper, we have presented an event-based framework for characterizing the evolution of dynamic interaction graphs. The framework is based on the use of certain critical events that facilitate our ability to compute and reason about novel behavior-oriented measures, which can offer new and interesting insights for the characterization of dynamic behavior of such interaction graphs. We have demonstrated how measures for Sociability, Stability, Influence and Popularity can be compiled. Note, that these are by no means exhaustive. The advantage of our general event-detection framework is that it can be used to derive other types of custom behavioral measures as well, which is extremely useful in the context of social information management.

Our framework does not make assumptions regarding snapshot lengths or the clustering algorithm used to identify clusters. Although, our scheme operates independently of the clustering algorithm chosen, we acknowledge the fact that the optimality of the clusters will play a part in the efficacy of the results obtained. We have shown in previous work on graphs in the proteomics domain [Asur et al. 2007] that ensemble clustering can be employed to improve the quality of clusters. Such methods can be applied to obtain efficient and robust snapshot clusters for the event-based framework. Also, in this work, we have relied on domain knowledge

to determine interval lengths for snapshots. In practice, in the absence of domain information, methods such as time series segmentation and smoothing can be used to derive suitable time intervals [Tadepalli et al. 2008].

We have shown how semantic content and category hierarchy information can be incorporated to reason about community-based events such as Merges and Splits. We have presented a diffusion model for evolving networks and have shown the use of behavioral patterns for influence maximization.

We have demonstrated the efficacy of our framework in characterizing and reasoning on three different datasets - DBLP, Wikipedia and a clinical trials dataset. The application of the behavioral patterns we obtained to a cluster link prediction scenario provided favorable results, with the Sociability Index producing a large number of accurate predictions. In particular, our experiments demonstrated that temporal change information can be extremely informative and can be well utilized for predictions in a broad range of applications. Apart from its general efficiency, the framework is also scalable and we have incorporated it in a visual toolkit for dynamic graphs [Yang et al. 2008].

In future work, we would like to extend the temporal analysis to graph production rules and graph grammars. Of particular interest in our context will be to evaluate if graph grammars can be inferred from such interaction networks by learning the *production rules* that govern evolution of clusters or neighborhoods. We would also like to extend our framework to reason and infer other behavioral patterns, as well as include other types of interaction graphs. Also, we would like to evaluate the usability of our framework in the verification of popular hypotheses in social network analysis [Palla et al. 2007; Leskovec et al. 2005].

8. ACKNOWLEDGEMENTS

This work is supported in part by the DOE Early Career Principal Investigator Award No. DE-FG02-04ER25611, NSF CAREER Grant IIS-0347662 and NSF SGER Grant IIS-0742999. We would like to thank Klaus Berberich and Evgeniy Gabrilovich for providing us with the Wikipedia dataset and the category hierarchy respectively. The authors would also like to thank Sameep Mehta for his useful comments and suggestions, and Xintian Yang for help in processing the Wikipedia dataset. We would also like to thank the anonymous reviewers who identified areas of our manuscript that needed modification and provided valuable comments in this regard.

REFERENCES

- ALKEMADE, F. AND CASTALDI, C. 2005. Strategies for the diffusion of innovations on social networks. *Computational Economics* 25, 1-2.
- ASUR, S., PARTHASARATHY, S., AND UCAR, D. 2007. An event-based framework for characterizing the evolution of interaction graphs. *Technical Report Feb 2007, OSU-CISRC-2/07-TR16.*
- ASUR, S., UCAR, D., AND PARTHASARATHY, S. 2007. An ensemble framework for clustering protein protein interaction networks. *Bioinformatics* 23, 13, i29-40.
- BACKSTROM, L., HUTTENLOCHER, D. P., AND KLEINBERG, J. M. 2006. Group formation in large social networks: membership, growth, and evolution. *SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- BARABASI, A.-L. AND BONABEAU, E. 2003. Scale-free networks. *Scientific American* 288, 60-69.
- ACM Transactions on Computational Logic, Vol. 2, No. 3, 09 2001.

- BARABASI, A.-L., JEONG, H., RAVASZ, R., NDA, Z., VICSEK, T., AND SCHUBERT, A. 2002. On the topology of the scientific collaboration networks. *Physica A* 311, 590–614.
- CHAKRABARTI, D., KUMAR, R., AND TOMKINS, A. 2006. Evolutionary clustering. *SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- CHI, Y., SONG, X., ZHOU, D., HINO, K., AND TSENG, B. L. 2007. Evolutionary spectral clustering by incorporating temporal smoothness. *SIGKDD International Conference on Knowledge Discovery and Data Mining*, 153–162.
- CLAUSET, A., NEWMAN, M. E. J., AND MOORE, C. 2004. Finding community structure in very large networks. *Physical Review E* 70.
- COWAN, R. AND JONARD, N. 2004. Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control* 28, 1557–1575.
- FALKOWSKI, T., BARTELHEIMER, J., AND SPILIOPOULOU, M. 2006. Mining and visualizing the evolution of subgroups in social networks. *IEEE/WIC/ACM International Conference on Web Intelligence*.
- FERLEZ, J., FALOUTSOS, C., LESKOVEC, J., MLADENIC, D., AND GROBELNIK, M. 2008. Monitoring network evolution using mdl. *IEEE International Conference on Data Engineering (ICDE)*, 1328–1330.
- FLAKE, G. W., LAWRENCE, S. R., GILES, C. L., AND COETZEE, F. M. 2002. Self-organization and identification of web communities. *IEEE Computer* 36, 66–71.
- GABRILOVICH, E. AND MARKOVITCH, S. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. *International Joint Conference on Artificial Intelligence (IJCAI)*.
- GANESAN, P., GARCIA-MOLINA, H., AND WIDOM, J. 2003. Exploiting hierarchical domain structure to compute similarity. *ACM Trans. Inf. Syst* 21, 1.
- GIRVAN, M. AND NEWMAN, M. E. J. 2002. Community structure in social and biological networks. *Proc. National Academy of Sciences of the United States of America* 99, 12, 7821–7826.
- HOPCROFT, J., KHAN, O., KULIS, B., AND SELMAN, B. 2004. Tracking evolving communities in large linked networks. *Proceedings of the National Academy of Sciences* 101, 5249–5253.
- KEMPE, D., KLEINBERG, J., AND TARDOS, E. 2003. Maximizing the spread of influence through a social network. *SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- KEMPE, D., KLEINBERG, J., AND TARDOS, E. 2005. Influential nodes in a diffusion model for social networks. *Proc. Intl. Colloquium on Automata, Languages and Programming (ICALP)*.
- LADA, A. A., , AND ADAR, E. 2003. Friends and neighbors on the web. *Social Networks* 25, 3 (July), 211–230.
- LESKOVEC, J., KLEINBERG, J. M., AND FALOUTSOS, C. 2005. Graphs over time: densification laws, shrinking diameters and possible explanations. *SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- LESKOVEC, J., LANG, K. J., DASGUPTA, A., AND MAHONEY, M. W. 2008. Statistical properties of community structure in large social and information networks. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*. ACM, New York, NY, USA, 695–704.
- LIBEN-NOWELL, D. AND KLEINBERG, J. M. 2003. The link prediction problem for social networks. *Proc. ACM CIKM Intl. Conf. on Information and Knowledge Management*.
- LIN, D. 1998. An information-theoretic definition of similarity. *Proc. 15th Intl. Conf. Machine Learning*.
- LORD, P., STEVENS, R., BRASS, A., AND GOBLE, C. 2003. Semantic similarity measures as tools for exploring the gene ontology. In *Pacific Symposium on Biocomputing*, 601–612.
- NEWMAN, M. 2001. Clustering and preferential attachment in growing networks. *Phys. Rev. E* 64.
- NEWMAN, M. E. J. 2006. Modularity and community structure in networks. *Proc. National Academy of Sciences of the United States of America* 103, 23, 8577–8582.
- OTEY, M. E., PARTHASARATHY, S., AND TROST, D. C. 2006. Dissimilarity measures for detecting hepatotoxicity in clinical trial data. *SIAM International Conference on Data Mining (SDM)*.
- PALLA, G., BARABASI, A.-L., AND VICSEK, T. 2007. Quantifying social group evolution. *Nature* 446, 7136 (April), 664–667.
- ACM Transactions on Computational Logic, Vol. 2, No. 3, 09 2001.

- RESNIK, P. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* 11, 95–130.
- RICHARDSON, R., SMEATON, A. F., AND MURPHY, J. 1994. Using WordNet as a knowledge base for measuring semantic similarity between words. *Artificial Intelligence and Cognitive Science (AICS)*.
- SAMTANEY, R., SILVER, D., ZABUSKY, N., AND CAO, J. 1994. Visualizing features and tracking their evolution. *IEEE Computer* 27, 7, 20–27.
- STREHL, A. AND GHOSH, J. 2002. Cluster ensembles a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research* 3, 583–617.
- TADEPALLI, S., RAMAKRISHNAN, N., WATSON, L. T., MISHRA, B., AND HELM, R. F. 2008. Simultaneously segmenting multiple gene expression courses by analyzing cluster dynamics. *Asia Pacific Bioinformatics Conference (APBC)*, 297–306.
- TANTIPATHANANANDH, C., BERGER-WOLF, T. Y., AND KEMPE, D. 2007. A framework for community identification in dynamic social networks. *SIGKDD International Conference on Knowledge Discovery and Data Mining*, 717–726.
- WASSERMAN, S. AND FAUST, K. 1994. Social network analysis. *Cambridge University Press*.
- YANG, H., PARTHASARATHY, S., AND MEHTA, S. 2005. A generalized framework for mining spatio-temporal patterns in scientific data. *SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- YANG, X., ASUR, S., PARTHASARATHY, S., AND MEHTA, S. 2008. A visual-analytic toolkit for dynamic interaction graphs. *SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Received December 2007; Revised September 2008; accepted December 2008