IBM Research Report

On Quantifying Changes in Temporally Evolving Dataset

Rohan Choudhary
Dept. of Computer Science and Engineering
Indian Institute of Technology
New Delhi - 110016. India.

Sameep Mehta
IBM Research Division
IBM India Research Lab
Plot No -4 Phase 2 Block-C, Vasant Kunj
New Delhi - 110070, India
Email: *sameepmehta@in.ibm.com*

Amitabha Bagchi
Dept. of Computer Science and Engineering
Indian Institute of Technology
New Delhi - 110016. India.

IBM Research Division
Almaden - Austin - Beijing - Delhi - Haifa - T.J. Watson - Tokyo - Zurich

Abstract

In this paper, we present a general framework to quantify changes in temporally evolving data. We focus on changes that materialize due to evolution and interactions of features extracted from the data. The changes are captured by the following key transformations: *create, merge, split, continue,* and *cease*. These transformations have been used successfully to study the temporal dimension in various domains like text, scientific data, bioinformatics and social networks. Once discovered, the user has to sift through the discovered transformations to find useful ones. This is a cumbersome and error prone approach. Moreover, the number of discovered events can be large (e.g. 100000 events in DBLP) which renders the manual process infeasible. Towards this goal, we present a general algorithm to automatically rank/score these transformation. First, we identify various factors which influence the importance of each transformation. These factors are then combined using a weight vector. The weight vector encapsulates domain knowledge. We evaluated our algorithms on various real data sets arising from diverse domains. In this paper, we present our results from the following datsets: *DBLP, IMDB, Text and Scientific Dataset.*

1 Introduction

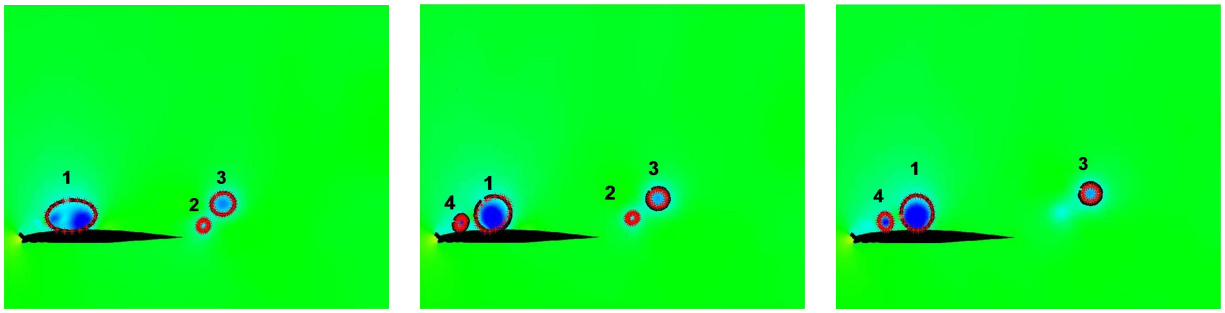
In this paper we propose methods to quantify the changes in a temporally evolving dataset. These dynamic datasets are studied to understand the evolution of a network by studying various properties like diameter [11]. In recent times, these datasets have also been studied to discover "important" phenomena, features and time steps [4]. Furthermore, such an analysis also explains how an individual feature evolves over time and how multiple features interact with each other [19]. This analysis, if performed in a proper fashion, provides tremendous insights into the data and the associated underlying physical processes. To study the changes in such datasets, Wang and Silver [16] proposed a set of five critical events *create,continue, split, merge* and *death* to capture interaction and evolution of features. We explain these events in more details in next section. These events have been used in various domains like scientific data-including datasets from computational fluid dynamics [18], computational molecular dynamics, bioinformatics [12] , social networks [4], text (news corpora) [6] and time varying clusters [17].

Most of the previous work has focused on developing algorithms for efficiently mining the critical events and present them to the user. The user has to manually examine each event and find important/useful ones. This manual process is cumbersome and error prone. Moreover, the sheer number of events that can be potentially generated from the dataset makes this manual process improbable (if not impossible). For example, in DBLP dataset, more than 100,000 such events were discovered. Manual inspection of all these events will render the whole exercise useless. There is no algorithm to automatically quantify or rank the events generated from different domains. Asur et al [4] presented some measures to capture sociability and influence of nodes in dynamic social networks. Choudhary et al [6] presented mechanism to capture interactions among important actors in news corpora. However, these efforts were domain specific. In the paper we present general algorithms which can be applied to datasets generated from various domains. The algorithms takes into account properties of individual features as well as the interaction among the features.

To develop our algorithm, we first study those properties of a feature which influence its importance. For example, if a cluster corresponds to a feature, number of members in a cluster (size) play an important role in defining its importance. It is clear that all the properties are not valid/useful in all domains. For example, size of an individual actor is not relevant in the context of the IMDB data. Therefore, in a given domain, the relative importance of the properties are captured using a weight vector. This vector enables the user to embed domain expertise in the algorithm. We discuss these properties, key rationale behind including them and weight vectors in detail in subsequent sections.

We evaluated the proposed algorithm on the following datasets *DBLP_author*, *DBLP_cluster, IMDB, Text* and *Scientific Dataset*. We present two different views *Author* and *Cluster* of DBLP dataset. Different features are extracted in each view which enables us to showcase the generality of the proposed algorithm. We use news corpus to demonstrate the applicability of algorithms on text. For scientific dataset we use Computational Fluid Dynamics (CFD) simulation datasets. The IMDB dataset is used to highlight the efficiency of the algorithms. The dataset consist of *1,000,000* features (nodes) with *34,000,000* interactions (edges). The algorithm takes *tt* minutes to discover all events and *tt* minutes to assign a score to each event. The details of dataset and associated feature extraction mechanism are presented in Section 4

The rest of the paper is organized as follows: Section 2 explains critical events with expository examples. The section also describes the past research which is most pertinent to this work. The



T1 T2 T3

Figure 1:

proposed scoring scheme is presented in Section 3. Experimental setting and results are presented in Section 4. We conclude in section 6 and also highlight some of the ongoing and future initiative for this project in Section 5

2 Background and Related Work

In this section we describe critical events in details and also highlight exisiting research efforts which are closely related to the proposed work.

Wang and Silver [16] proposed a set of key transformations to capture evolution and interactions of features in time varying fluid dynamics datasets. These transformations are often referred to as *critical events*. Assume f_i^t represents i^{th} feature extacted at t^{th} time step/slice. The critical events are:

1. **Creation:** A creation event is identified if a new feature is detected at $t+1^{th}$ time step, i.e, $\neg\exists f_i^t \wedge \exists f_i^{t+1}$. The creation event is represented as $\boldsymbol{Create}^{\boldsymbol{f_i^{t+1}}}$.

2. **Continuation:** A continuation event is identified if a feature is detected at both t^{th} and $t+1^{th}$ time step, i.e, $\exists f_i^t \wedge \exists f_i^{t+1}$. The event is represented as $\boldsymbol{Continue}_{\boldsymbol{f_i^t}}^{\boldsymbol{f_i^{t+1}}}$.

3. **Dissipation:** A dissipation event is identified if an existing feature is not detected at $t+1^{th}$ time step, i.e, $\exists f_i^t \wedge \neg\exists f_i^{t+1}$. A dissipation event is represented as $\boldsymbol{Dissipate}_{\boldsymbol{f_i^t}}$.

4. **Merge:** A merge event is identified when two or more features at t^{th} time step join to form a larger feature at $t+1^{th}$ step. The conditions for a valid merge are:

$$\exists f_i^t \wedge \exists f_k^t \wedge \neg\exists f_l^t \tag{1}$$

$$\neg\exists f_i^{t+1} \wedge \neg\exists f_k^{t+1} \wedge \exists f_l^{t+1} \tag{2}$$

   The merge event is represented as $\boldsymbol{Merge}_{\boldsymbol{f_i^t},\boldsymbol{f_k^t}}^{\boldsymbol{f_l^{t+1}}}$

5. **Split:** A split event is identified when a single feature at t^{th} time step breaks to form two or more features at $t+1^{th}$ step. The conditions for a valid split are:

$$\neg\exists f_i^t \wedge \neg\exists f_k^t \wedge \exists f_l^t \tag{3}$$

$$\exists f_i^{t+1} \wedge \exists f_k^{t+1} \wedge \neg\exists f_l^{t+1} \tag{4}$$

   The split event is represented as $\boldsymbol{Split}_{\boldsymbol{f_l^t}}^{\boldsymbol{f_i^{t+1}},\boldsymbol{f_k^{t+1}}}$

Please note the relationship between creation and dissipation events. One changes into another if time dimension is reversed. Similar argument holds for merge and split transformation. We use this property for expository purposes. Figure 1 shows the key transformations discovered in computational fluid dynamics dataset. The data simulates airplane wing in 3D and a cross section (region in black) is taken to generate 2D data. The three images are three different time steps (not in any particular order). The features or regions of interest (ROI) in this dataset are vortices. The

vortices at each time step are identified using algorithms presented by Jiang et al. [8]. The vortices are approximated by ellipses which are shown in the images. Each vortex is assigned a unique numeric label. From T1 to T2 vortex 1 *splits* into vortex 1 and 4. Similarly, vortex 2 *dissiapated* from (T1)T2 to T3 whereas vortex 3 *continues* in all three time steps. Going from T2 to T1, vortex 1 and 4 *merge* to form larger vortex 1. Similarly, vortex 2 is *created* from T3 to T1 and T2.

Our work build on seminal work by Silver and Wang [16]. The authors proposed the above-mentioned critical events. The critical events framework has been extended for various domains including bioinformatics, clustering etc. Recently, Asur et al [4] presented extended these ideas for dynamic graphs. The authors also proposed measures to capture some properties like socialbility of nodes. Choudhary et al [6] employed the events mining algorithm for news corpus and also presented a ranking mechanism. However, both these efforts were domain specific and the proposed algorithm cannot be directly applied to other domains. EventRank [15] is a system which takes into account the sequence of actions to assign rank to nodes in time varying data. The algorithm extends HITS/PageRank to include temporal information. The goal and approach of ranking in [15] is different from the one presented in this paper. Visualization has also been used to explore dynamic graphs to discover important nodes and phenomena. Kumar and Garland [10] presented algorithms to visualizse time varying graphs. The article focus on graph layout, stratification and rendering issues. Recently, we [20] presented a visual toolkit to explore the dynamic social networks. The toolkit provides several filters to the user for exploration. Simple measure like socialibility and influce are used to provide some ranking to individual nodes. However, quantifying the changes in time varying graph was not considered in any of above mentioned efforts. Moreover, these efforts were tailored for dynamic graphs not for general time varying datasets.

3 Algorithm

In this section, we describe our algorithm. First we present the notations used in this section. Next we present the key properties of the feature which should be taken into account while assigning a score. Finally, we present the scoring algorithm for each of the transformation.

Basic Notation and Explanation

$\mathcal{S} = (S_1, S_2, \ldots, S_T)$ denotes a time varying data with T time steps. f_k^t denotes k^{th} feature extracted from t^{th} time step. Wherever unambiguous we omit subscript. $N_{[t_1,t_2]}(f_k, \ldots, f_l)$ captures number of times the group by features $(f_k, \ldots, f_l)$ co-occurred during a given time interval $[t_1, t_2]$.

Feature Properties

We capture four properties of each feature. Collectively, these properties capture information about evolution of a single feature and its interaction with others. We consider the following properties:

1. $\boldsymbol{Strength_{t_1}^{t_2}}$ Strength captures importance of a feature vis-a-vis the whole dataset. Informally, strength of f compares occurrence of F with the most frequent feature. Mathematically, strength is defined as:-

$$Strength_{t_1}^{t_2}(f) = \frac{N_{t_1}^{t_2}(f)}{max_f(N_{t_1}^{t_2}(f))} \tag{5}$$

   The definition of strength can be easily extended to compute average, maximum and minimum strength of a group of features. In some cases the group strength is more relevant than individual strengths of members of the group. For example, in case of clusters, the strength of a cluster is better defined as the average of the individual strength of its members. Similarly, in the case of news, when a feature consists of two actors, the strength of the feature is better defined as the average strength of the two actors.

2. $\boldsymbol{Coupling_{t_1}^{t_2}(f_k, f_l)}$ Coupling captures the interaction of the features within a time period. The interaction we are interested is co-occurrence. There are multiple ways to define the coupling. One way is to calculate the fraction of time steps in which f_k and f_l co-occurred ,i.e, $\frac{N_{[t_1,t_2]}(f_k,f_l)}{t_2-t_1}$. However, this metric is negatively biased towards less frequent features. For example, if f_k and f_l always co-occurred but occurred only twice in 100 time steps the coupling is $\frac{2}{100}$. While the features always interacted the coupling is low. To handle this problem we use the following definition to calculate coupling:

$$Coupling_{t_1}^{t_2}(f_k, f_l) = \frac{N_{t_1}^{t_2}(f_k, f_l)}{N_{t_1}^{t_2}(f_k) + N_{t_1}^{t_2}(f_l) - N_{t_1}^{t_2}(f_k, f_l)} \tag{6}$$

Event	Score
$Continue^{f_k^{t+1}}_{f_k^t}$	$w_1 * Size(f_t) + w_2 * Strength^{t+1+F}_{t+1-P}(f_k) + w_3 * GCoupling^{t+1+F}_{t+1-P}(f_k, f_k)$
$Creation^{f_k^{t+1}}$	$w_1 * Size(f_k^{t+1}) + w_2 * Strength^{t+1+F}_{t}(f_k) + w_3 * GCoupling^{t+1+F}_{t}(f_k, f_k)$
$Dissipate_{f_k^t}$	$w_1 * Size(f_k^{t}) + w_2 * Strength^{t}_{t-P}(f_k) + w_3 * GCoupling^{t}_{t-P}(f_k, f_k)$
$Merge^{f_m^{t+1}}_{f_k^t, f_l^t}$	$\frac{w_1 * (Size(f_k^t) + Size(f_l^t))}{2} + \frac{w_2 * (Strength^t_{t-P}(f_k) + Strength^t_{t-P}(f_l))}{2} + w_3 * GCoupling^{t+F}_t(f_m, f_m) + w_4 * Strength^{t+F}_t(f_m) - w_5 * Coupling^t_{t-P}(f_k, f_l)$
$Split^{f_k^{t+1}, f_l^{t+1}}_{f_m^t}$	$\frac{w_1 * (Size(f_k^{t+1}) + Size(f_l^{t+1}))}{2} + \frac{w_2 * (Strength^{t+F}_t(f_k + Strength^{t+F}_t(f_l))}{2} + w_3 * GCoupling^t_{t-P}(f_m) + w_4 * Strength^t_{t-P}(f_m) - w_5 * Coupling^{t+F}_t(f_k, f_l)$

Table 1: Score for each Key Event

With this definition we will get coupling of 1 for the above example. An interesting case to consider is if f_k occurs 2 times but f_l occurs 100 times in 100 time steps, the coupling is $\frac{2}{100}$. One can argue that since whenever f_k has occurred it has interacted with f_l, therefore the coupling (w.r.t. to f_k) should be high. However, we are interested in joint interaction between f_k and f_l not conditional coupling.

Similar to the strength, coupling can be extended for a group of features. Coupling of two groups $\mathcal{G}_1$ and $\mathcal{G}_2$ is given as:

$$GCoupling^{t_2}_{t_1}(\mathcal{G}_1, \mathcal{G}_2) = \frac{1}{N}(\sum_{f_k \in \mathcal{G}_1} \sum_{f_l \in \mathcal{G}_2} Coupling^{t_2}_{t_1}(f_k, f_l)) \tag{7}$$

where $N \leq |\mathcal{G}_1| * |\mathcal{G}_2|$ is the averaging factor, i.e., the number of pairs which have non-zero coupling. The group coupling can be used to compute coupling between two clusters. In such cases, coupling will capture the inter-cluster member interactions. Moreover, **self coupling**, i.e, $GCoupling^{t_2}_{t_1}(\mathcal{G}_1, \mathcal{G}_1)$ captures coherence of a group. For example, a cluster with very high self Coupling implies that the cluster members have high interaction. Please remember that the interaction here captures cooccurences. Therefore, we can infer that the cluster is a stable one with few changes in its members.

3. ***Size(f_k)*** Size of the feature is self explanatory. For some features like IMDB actors, size is not relevant. Whereas in other domains like CFD or clustering applications, size is extremely important. The exact calculation of size is domain Dependant. In this paper size of a group/cluster is defined as number of members in that particular group. For vortices size is calculated as area the enclosing ellipse.

Please note that each of the properties above are normalized between [01]. We are now in a position to develop metrics to quantify the key transformations. As mentioned earlier, all the above described properties are not relevant in the context of every dataset. Moreover, some properties may be more important than others. To handle these cases an importance/weight vector is used. The vector captures relative importance of different properties and other domain specific knowledge. Section 4 presents the actual weight vector used to generate the results in this paper. Next, we present quantifying metrics for each transformation along key rationale behind the metric.

1. ***Continue*** $^{f_k^{t+1}}_{f_k^t}$ The score of a continue of a feature is given in Table 1 Row 1. The overall score is composed of three terms. The first part accounts for the size of the feature. The second terms captures the importance (frequency of occurrence) of the feature in its temporal neighborhood. Finally, the last term measure the stability of the feature in the temporal neighborhood. The idea of using a neighborhood is motivated by the fact that in most systems the features are influenced by their immediate neighbors. Far-off neighbor have little or no

impact. This property is extensively used in sptio-temporal application, pre-fetching in cache and markov models. One key difference is that our temporal neighborhood also looks ahead. This is because if a feature is not useful in the future then its current importance should reflect this. For example consider two features F_1 and F_2 which are exactly same. Both F_1 and F_2 continue from $t = 8$ to $t = 9$ however F_1 dissipates at $t = 10$ whereas F_2 continues till $t = 15$. In such cases without looking ahead both will be assigned same score which is not desirable. Therefore, we consider both recent part and upcoming future to assign the score. Informally, a continue is important if *a frequent and stable feature of considerable size continues*. Finally, since $\sum w_i = 1$, the continue score lies between 0 and 1. We show how the actual values of weight vector for each domain in Section 4

2. $\boldsymbol{Creation}^{f_k^{t+1}}$ The score of a birth of a feature is shown in Table 1 Row 2. Informally, a creation is important *if the new feature is large and will be frequent and stable in near future*. The metric is very similar to that of continue transformation. The only difference is that the temporal neighborhood only considers future time steps because the feature is non-existent earlier.

3. $\boldsymbol{Dissipate}_{f_k^t}$ The score of a death of a feature is demonstrated in Table 1 Row 3. The scoring mechanism very similar to that of creation event, except that the time period considered is before the death event, for obvious reasons. A dissipation is important*if a large, frequent and stable feature dissipates*. The creation and the dissipation scores also lie between 0 and 1.

4. $\boldsymbol{Merge}^{f_m^{t+1}}_{f_k^t, f_l^t}$: Using the properties described above the score of a merge event shown in Table 1 Row 4. The merge event scoring is more intricate. We again explain each of term in the function. The first and second term captures the average size and strength of merged features because typically merging of two large (frequent) features is more interesting than merging of smaller (infrequent) features. The third and forth term captures strength and stability of newly created feature in future to ascertain that the new feature is indeed important. The last term captures the past interaction of the merged features. High value of this term has negative impact on the overall merge score. This is because if the two merged features were also interacting in the past than the new merge is not very interesting. On the other hand, if two features with no past interaction merge then it may point to an interesting underlying phenomenon which warrants more attention and hence this merge should be rated higher than the previous one. A merge is important if *two large, frequent and non-interacting features merge (interact) to form a frequent and stable feature*. Please note that the size of new feature will be approximately equal to the sum of sizes of merged features[1]. Therefore, we don't explicitly consider size of the new feature. The merge score can lie between -1 and 1.

5. $\boldsymbol{Split}^{f_k^{j+1}, f_l^{j+1}}_{f_m^j}$ Table 1 Row 2 specifies the score function of a split event. The metric is very similar to that of a merge because of the converse nature of these two transformations. The last term here implies that the new features should have less or no interaction for the split to be interesting. A split is important if *a frequent and stable features splits to form two large, frequent and non-interacting features*. The split score can lie between -1 and 1.

Order Complexity: For each discovered transformation, we use information from the temporal neighbourhood to rank the transformation. Let the size of the temporal neighbourhood be H timestamps. Then the time complexity to rank one transformation is $O(H)$. If F donotes the number of features at each time step then F transformations are extracted (one for each feature) and ranked. The complexity for each time step is $O(H * F)$. If there are T time stamps, the total complexity becomes $O(H * F * T)$ where H < T.

4 Experiments

In this section we present the detailed evaluation of our algorithm on various datasets. First, we provide details about the datasets and the definition of features in each of the dataset. Next, we present a discussion on the weight vector and finally we present ranked events mined from the dataset.

[1]This is true for majority of domains. However, if this property doesn't hold the size of new feature can be included as a term in the metric

Dataset	Time Steps	Merges	Splits	Deaths	Creations	Continues
Scientific Dataset	5000	32	22	32	46	1278
DBLP Pair Based	31	138281	169939	267487	366072	217975
IMDB Pair Based	40	799920	901932	237443	261392	71479
DBLP Cluster Based	31	1157	1255	137230	144318	5163
IMDB Cluster Based	40	79	80	13675	15008	40

Table 2: Dataset Description and Number of Discovered Transformations

	Merge——					Split——					Others		
Dataset	w_1	w_2	w_3	w_4	w_5	w_1	w_2	w_3	w_4	w_5	w_1	w_2	w_3
News Corpus	0	0.33	0.33	0	0.33	0	0	0.33	0.33	0.33	0	1	0
DBLP Pair Based	0	0.25	0.25	0.25	0.25	0	0	0.33	0.33	0.33	0	1	0
IMDB Pair Based	0	0	0.33	0.33	0.33	0	0.25	0.25	0.25	0.25	0	1	0
DBLP/IMDB Cluster	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.33	0.33	0.33
Scientific Data	0.50	0	0	0.50	0	0.50	0	0	0.50	0	0.50	0.50	0

Table 3: Weights Vector for Different Datasets

4.1 Dataset Description

- **Text Data-** The text corpus we used for our evaluation purposes consists of news stories related to the news topic of US Elections 2004.The dataset was mined from http://www.nytimes.com and contains 389 news stories published between February 02, 2004 and November 25, 2004. Each story can have multiple actors. These actors are extracted using the algorithms presented by Mei and Zhang [13, 14]. These time stamped actors form the temporally evolving data in this case. More details about this data can be found in Choudhary et al [6].

- **Scientific Dataset-** The scientific dataset is generated using computational fluid dynamics (CFD) simulations. The data presents the simulation of an airplane wing's cross section. The data consists of velocity vector computed at each grid point. The data is generated for 5000 time steps. Vortices are features in this dataset. We use algorithms proposed by Jiang et al [8] to detect and extract vortices at each time step. Key events are mined using method outlined by Silver and Wang [16].

- **DBLP Dataset-** The DBLP data is downloaded from `www.informatik.uni-trier.de/ ley/db`. The data is in the form of a XML file containing information about *Title of Paper/Publication, Author Names, Year of Publication* and *Conference/Journal/Book* where it was published. We processed DBLP data from the year 1975 to 2006 (31 years). The number of authors listed in DBLP till 2006 were found to be around 0.5 million and the number of publications between them were around 0.9 million. Each author is represented as a node and collaboration between authors is captured by an edge. Such a graph is generated for each year. We generated two different views from this data. **Author View:** In this view, authors are the features and we capture interactions among two authors. **Cluster View:** We follow the methodology proposed by Asur et al. [5] to generate this view. Each graph is partitioned using METIS [9]. The subgraphs are the features and transformations capture interactions among the subgraphs.

- **IMDB Dataset-** Another dataset chosen to evaluate our scoring algorithms is the Internet Movie Database (IMDB). A publicly available dataset [3] was downloaded from the IMDB's ftp site and was parsed into a relational database. A subset of the IMDB dataset was chosen for evaluation purposes. We worked with the data of 40 years from 1960 to 2000 (40 years). Also, we filtered the data to contain only those movies which were filmed in USA. This was done, since IMDB contains data from film production industries around the world, and the data is too heterogeneous to be informative for analysis. The actors form the node and co-actors relationship is captured by an edge. We again generated the same two views as for DBLP dataset. The number of actors from 1960 to 2000 were in the order of 800,000 and the number of collaboration edges were in the order of 30 million.

We mined all the transformations from each dataset. Table 2 provides a description of the number of transformations mined from each dataset. We observe that the cluster based analysis of the IMDB dataset, yields very less transformations than its counterpart in DBLP, even though the size of IMDB is an order larger than that of DBLP. This is attributed to the fact that in the movie setting, it is highly unlikely that a cluster(group of actors) will act together in movies for consecutive years. Whereas in DBLP, we discover a group of researchers publish together for consecutive years

Date	Story and Creation of Actor
30/08	Republican Convention kicks off (convention)
13/04	Iraq issue starts coming up (Iraq)
06/07	Kerry chooses Edwards as running mate (Edwards, running mate)
14/05	Issue of same-sex marriage (same-sex marriage)
28/07	Issue of economy during democratic convention (economy)
28/07	Issue of global terrorism at democratic convention (terror)
01/08	Republicans challenge Kerrys Vietnam records (Vietnam)
13/05	Ralph Nader wins endorsement of Reforms Party (Nader)

Table 4: Synopsis of the top ranked creations in US Election 2004 corpus

Date	Story
03/03	Edwards bows out of Presidential Race (Edwards)
04/03	End of primaries (primaries)
15/10	End of third debate session (debates)
23/02	Howard Dean bows out of Presidential Race (Dean)
05/09	End of republican convention (convention)
17/06	Ralph Nader excluded from Presidential Debates (Nader)

Table 5: Top ranked cessation in US Election 2004 corpus.

because the members of the cluster are part of same university or research group. Therefore, we decided not to further analyze the cluster based view of IMDB.

4.2 Setting the Weight Vector

The generic scoring methodology, described in the previous section has the advantage that the weights used in each metric can be modified to incorporate domain specific knowledge. Next, discuss the weights vectors for different datasets and for different transformations.

Table 3 shows the list of weights used in different datasets for different transformations. Please recall w_1 is associated with size, w_2 with strength/frequency, w_3 with stability of a features, w_4 with past/future strength of features for merge/split and w_5 captures interaction amongst the features. Please note that w_4 and w_5 are defined only for merge and splits. For setting the weights, we adopted a simple methodology. The weights of the irrelevant components are set to zero and equal weights are assigned to other components, s.t. the weight add up to one. For example, the weight of size component (w_1) of news actor, DBLP author and movie actor (pair based) is set to 0, however for cluster view of same datasets, the size of clusters is important and ia assigned appropriate weight. Similarly, for create/continue/dissipate events only the strength (w_2) plays a role for News and DBLP/IMDB pair based view. However, size (w_1) plays a role for scientific data and cluster based view. The cluster based view (w_3) also take into account the stability of created/continued/dissipated cluster. Finally, for merge/split involving cluster view all 5 components are relevant and therefore, the weights are split equally amongst them. Other weights are be explained in a similar fashion.

4.3 Ranked Transformations

Now, we present the ranked transformation for all the datasets.

News Corpora - US Elections 2004

Table 4 shows the abstract of new stories where top 8 creations occurred in this corpus. The actual actors are also noted in the table. We note that all the creations point to major events/issues in US elections. Moreover, most of the creations continued for considerable period. Table 5 shows top cessation of actors. All the events points to end of a major phase in the election camp gain. Similarly, table 6 shows the abstract of the top 10 merges (scores in decreasing order) and corresponding dates identified in this corpus. The actors involved in merge are noted at the end of each headline. We notice that most of the major merges involve *Kerry* and *Bush*. This is because their strength is very high throughout the topic and thus whenever there is less interaction between them, then a new merger is ranked very high.

Scientific Dataset

The top ranked transformations discovered in the scientific dataset are displayed in Tables 7, 8 and 9. To put the things into perspective, the largest area vortex in the dataset had area 40000 units and the longest living vortex had a life of around 2000 frames. We can see that the top ranked continues mostly consist of objects of significant area and long life with respect to the other objects in the dataset. We also took into account the ratio of the sizes of the vortices merging together, while scoring merges and splits. The effect is seen, in the top ranked merges and splits. Most of the entries consist of comparable size moderate area objects rather than containing insignificant merges, involving one or more extremely small vortex. For merges we also show the lifetime -period for

Date	Story
25/02	Kerry starts winning a sequence of primaries (Kerry, primaries)
01/03	Discussion on Same-Sex marriage in the Democrat primaries (Same-sex marriage and Primaries)
20/08	Kerrys response to Bush on Vietnam Issues (Kerry and Bush)
15/10	Abortion comes up as an issue in the Bush-Kerry debates (Abortion and Debates)
12/09	GOP draws criticism from Kerry on Arms Ban (Kerry and Bush)
08/03	Battle for Florida hots up (Kerry and Bush)
12/08	Bush mocks Kerry on Vietnam (Kerry and Bush)
22/05	Bush visits Louisiana , where Democrats are campaigning (Bush and Democrats)
29/05	Kerry doubles his attack on Bush for Iraq War (Kerry and Bush)
30/10	Ruling by two federal courts to GOP (courts and GOP)

Table 6: Top ranked merges in US Election 2004 Corpus

Frame Number	Life	Area
39690	1340	25333.50
39450	480	39650
40280	1320	10740.5
38740	400	22907
39160	190	20029

Table 7: Top Continues in the scientific dataset

which the vortex continued after its creation. One can observe that vortices which were involved in high ranking merges have long lifetime which points that the merge event which lead to it's creation was indeed a important event. Similarly, the splits which are ranked higher involves one large vortex breaking into smaller but sizable vortices. We also show the lifetime of vortex which splits into smaller one. Again the vortices had large lifetime which implies that split (dissipation) was an important event.

DBLP Data : Cluster View

The top ranked creations in the DBLP cluster based analysis are displayed in Table 10. One of the top ranked creations in the DBLP dataset consisted of five authors Francesco Tortorella, Mario Vento, Pasquale Foggia, Carlo Sansone and Luigi P. Cordella, who published a paper together in 1997. After 1997, the five authors continued working at University of Naples and published number of papers in the field of machine learning and pattern recognition. Please note that a cluster will not be marked as continued if there exists even a single pair of authors who have co-authored a paper on one year but such a collaboration was missing in next year.

The top cessations in this domain are displayed in Table 11. The top ranked cessation group in the DBLP dataset consists of five researchers at the Mechanical Engineering Laboratory, AIST, Japan. The group published a number of papers in the area of self repairing systems from 1994 to 2003, after which they stopped publishing on this topic. This also marked the end of this collaboration, after which they started publishing outside this clique. This group was also ranked among the top 5 continues across the DBLP corpus. The high scoring Continue transformations are displayed in Table 12. Another interesting continue was of Flemming Nielson and Hanne Riis Neilson. They are married to each other and hardly published outside this collaboration from 1982 to 1996.

The top merges for the DBLP dataset are in Table 13. An interesting high ranking merge involved authors Richard Ostrowski, Bertrand Mazure, Lakhdar Sais and Eric Gregoire merging with Fred Hemery, Frederic Boussemart and Christophe Lecoutre in 2004. This merge is interesting because only Lakhdair Sais collaborated with the the other group, and just because of sheer high volume of such publications, the merge is ranked so high.

The top splits for the DBLP dataset are in Table 14. One of the top ranked splits involved the splitting of Jerry Gao and Yasufomi Toshiyama from Pei Hsia and David Chenho Kung in 1999.

Frame Number	Life	Area1	Area2	Area3
37590	180	3100	3754	7150
38330	110	1700	837	2565
39830	170	2828	1056	3616
37370	220	1765	546	2115
40700	30	511	126	601

Table 8: Top Merges in the scientific dataset

Frame Number	Life	Area1	Area2	Area3
39670	100	31264	5083	25678
39810	100	2793	1431	1056
40510	150	1852	612	1061
39710	40	5336	2954	2282
40680	30	668	126	101

Table 9: Top Splits in the scientific dataset

Author Names	Year of Creation
Kazuaki Ishizaki, Takeshi Ogasawara, Toshiaki Yasue, Toshio Suganuma, Tamiya Onodera, Mikio Takeuchi .. Motohiro Kawahito, Hideaki Komatsu, Toshio Nakatani, Kiyokuni Kawachiya	1999
Francesco Tortorella, Mario Vento, Pasquale Foggia, Carlo Sansone, Luigi P. Cordella	1997
Grigori Sidorov, Adolfo Guzman-Arenas, Igor A. Bolshakov, Sofia N. Galicia-Haro .. Alexander F. Gelbukh, Manuel Montes-y-Gomez, Aurelio Lopez-Lopez	1997
Stanley Y. W. Su, Herman Lam, Minsoo Lee, Joachim Hammer	1998
Janos Komlos, Endre Szemeredi, Miklos Ajtai	1980

Table 10: Top Cluster Creations in the DBLP dataset

Author Names	Year of Cessation
Satoshi Murata, Akiya Kamimura, Haruhisa Kurokawa, Eiichi Yoshida, Shigeru Kokaji, Kohji Tomita	2003
Mark M. Gourary, Sergey G. Rusakov, Sergey L. Ulyanov, Michael M. Zharov, Brian J. Mulvaney	2006
Henri Cohen, Francisco Diaz y Diaz, Michel Olivier	2003
Horacio M. Gonzalez Velasco, Carlos J. Garcia Orellana, Miguel Macias Macias, Ramon Gallardo Caballero	2005
Bjorg N. Cyvin, Sven J. Cyvin, Jon Brunvoll	1997

Table 11: Top Cluster Cessations in the DBLP dataset

Author Names	Duration of Collaboration
Antonella Santone, G. Vaglini, Roberto Barbuti, Nicoletta Francesco	1995-2007
Jean-Luc Hainaut, Jean Henrard, Jean-Marc Hick, Didier Roland, Vincent Englebert	1994-2005
Satoshi Murata, Haruhisa Kurokawa, Eiichi Yoshida, Shigeru Kokaji, Kohji Tomita	1998-2003
Fernando Cuartero, Fernando Pelayo, Valentin Ruiz, Diego Cazorla, Juan Pardo	2000-2007
Flemming Nielson, Hanne Riis Nielson	1982-1996

Table 12: Top Cluster Continuations in the DBLP dataset

Author Names	Year of Merge
Umberto Nanni(A), Marco Protasi (A), Alberto Spaccamela (A) Giorgio Gambosi (B), Guiseppe Italiano (B), Enrico Nardelli (B), Maurizio Talamo (B)	1989
Richard Ostrowski (A), Bertrand Mazure (A), Lakhdar Sais (A), Eric Gregoire (A) Fred Hemery (B), Frederic Boussemart(B), Christophe Lecoutre(B)	2004
Doru Tanasa (A), Brigitte Trousse (A) Maguelonne Teisseire (B), Pascal Poncelet (B), Florent Masseglia (B)	2003
Byoungro So (A), Mary Hall (A) Pedro Diniz (B), Joonseok Park(B)	2001
Nicola Ancona (A), Massimiliano Nittin (A), Ettore Stella (A), Anotenella Branca (A) Tiziana Orazio (B), Grazia Cicirelli (B)	2002

Table 13: Top Cluster Merges in the DBLP dataset

Author Names	Year of Split
David Chenho King (A), Pei Hsia (A), Chih-Tung Hsu (A) Yasufomi Toshiyama (B), Chris Chen (B), Jerry Gao (B)	1998
Zdenek Kouba (A), Tomas Vlcek (A) Vladimir Marik (B), Jiri Lazansky (B), Olga Stepankova (B)	1994
Jan Bergstra (A), Jan Willem Klop (A), J. V. Tucker (A) John Ch. Meyer (B), Jeffery Zucker (B), J.W. Bakker (B)	1984
John Shortle (A), Percy Brill (A) Martin Fischer (B), Denise Masi (B)	2004
Christine Jacquine (A), Laura Monceaux (A) Emmanuel Desmontils (A), Anne Vilnat (B), Anne Ligozat (B), Isabelle Robba	2005

Table 14: Top Cluster Splits in the DBLP dataset

Author Names	Duration of Collaboration
Andrzej Ehrenfeucht, Grzegorz Rozenberg	1973-2007
Alberto L. Sangiovanni-Vincentelli, Robert K. Brayton	1985-2008
Alfred V. Aho, Jeffrey D. Ullman	1968 - 1986
Irith Pomeranz, Sudhakar M. Reddy	1991 - 2008
Leonidas J. Guibas, Micha Sharir	1986 - 2000
Svetlana P. Kartashev, Steven I. Kartashev	1973-1987

Table 15: Top Author Pair Continuations in the DBLP dataset

Before this, the group published a large number of papers together. Around 1999, Jerry Gao moved to San Jose state university and Yasufomi Toshiyama opened a new company in San Jose, thus possibly reducing the collaboration possibilities, resulting in the discovered split.

DBLP Data : Pair View

The tables of top ranking transformations in DBLP pair dataset are Tables 15, 16 and 17. One of the top ranked pair continuations in the DBLP dataset involved Jeffrey D. Ullman and Alfred V. Aho, who collaborated together on significant works in databases and algorithms, during the 1970's.

The second ranked merge in the DBLP dataset corresponds to a single publication by authors Francky Catthoot and Mahmut Kandemir in 2003. Even though they published just a single paper together, they individually have published and contributed so much to their respective areas, that this merge was ranked as a significant one. **Please note that this merge is captured because we take into account strengths of individual features into account.**

Maer and Ullman figure both among the top ranked merges and splits, since they interacted very strongly only during 1979-1986 and were prominent researchers in their specific areas. Due to their impact on scientific community during and after their interactons, this split is ranked very high.

IMDB Data : Pair View

The top ranked actor Deaths for the IMDB Dataset are displayed in Table 18. We can see that the famous oscar nominated actors like James Mason and Denholm Elliott are part of this table. Other people who figure at the top include Cameron Mitchel(I), TV star during the 1960's. Interestingly, all the top ranked ceases actually coincided with the real life death of the actor involved.

The top actor births in the IMDB dataset are displayed in Table 19. Noteworthy top ranked people include stand up comics like Tom Kenny and stage artists like Ron Perlman. The biography on the IMDB website mentions that Perlman's career spanned three decades and he has worked with diverse actors from Marlon Brandon to Christina Ricci. Thus his ranking at the top is defendable in a scoring mechanism. Samuel Jackson is another award winning actor featuring in our list of top 5 Births because of the number of movies/shows he has appeared in.

The top 3 actor splits in the IMDB dataset are featured in Table 20. The top actor split includes puppeteers Jerry Nelson and Dave Goelson, who did a number of muppet movies together before 1985 and later Jerry Nelson became more focussed on his music career during 1990s, thus resulting in a split in 1985. Another very surprising feature in the top splits include former US president Lyndon Johnson and famous leader Martin Luther King. Both of them were featured in a number of documentaries during the 1990s. In the next decade, the fewer number of such documentaries

Author Names	Year of Merger	Duration of Collaboration
Jayadev Misra, K. Mani Chandy	1979	1979 - 1987
Francky Catthoor, Mahmut Kandemir	2003	2003 - 2003
Bernard Chazelle, Micha Sharir	1989	1989 - 1996
Jeffrey D. Ullman, David Maier	1979	1979 - 1986
Stephen L. Bloom, Zoltn sik	1988	1988 - 2008

Table 16: Top Author Pair Merges in the DBLP dataset

Author Names	Year of Split	Duration of Collaboration
Hermann A. Maurer, Derick Wood	1983	1976 - 1983
Eitan M. Gurari, Oscar H. Ibarra	1983	1978 - 1983
Noureddine Belkhatir, Walclio L. Melo	1994	1991 - 1994
Detmar W. Straub, Ephraim R. McLean	1997	1994 - 1997
David Maer, Jeffrey D.Ullman	1986	1979 - 1986

Table 17: Top Author Pair Splits in the DBLP dataset

Actor Names	Career Tenure
James Mason	1954-1984
Scatman Crothers	1951-1986
Derek Lyons	1977-1991
Denholm Elliott	1949 - 1992
Cameron Mitchel(I)	1945 - 1994

Table 18: Top Actor Deaths in the IMDB dataset

contributed to the split in 2000.

The top 5 author merges in the IMDB dataset are displayed in Table 21. Each of the result is an interesting event in a different way. The top ranked result is of professional wrestlers Hulk Hogan and Kevin Nash, who came together at the World Championship Wrestling forum. One of the high ranking results include the merge of actors Sid Caesar and Howard Morris in 1998. After 1998, Howard Morris featured in many documentaries and movies, paying tribute to Sid Caesar, thus resulting in a top ranked merge.

The top 5 actor continues in the IMDB dataset is displayed in Table 22. Each one of the entries, represents a successful relationship. Frank Welker, known as the voice god of hollywood and Michael Bell, another voice star, feature as the top ranked continuing relationship. The number of such relationships between voice stars, forms 80% of the top continues in the IMDB dataset, simply because the number of such professionals working in the area of dubbing and voice are less, and hence the number of movies common to a pair is very high. Jim Henson and Frank Oz, one of the most famous puppeteers in hollywood, form the second most continuing relationship. The relationship between Shawn Michaels, one of the longest continuing professional wrestler and Mark Yeato, the professional time keeper at the world wrestling entertainment forum, is a very interesting relationship and another one out of the set of many top ranked relationships formed from the wrestling industry.

5 Discussion

In this paper, we presented a generic framework to quantify changes in evolving data.We proposed an event ranking mechanism, in order to find the interesting and informative events inside such data. In future, we plan to extend the overall framework in multiple directions. Three particular extensions of this work seem very promising at present.

1. **Study the Implications of Temporal Neighbourhood for Streaming Data:** Please note that we don't consider streaming data in this paper. However, it is evident that the look ahead property of our temporal neighborhood will pose a problem in streaming setting. There are two ways to handle this. First, a buffer can be introduced which store F time step and the results will have a lag of F time units. In such scenario no change is needed in the above mentioned methodology. Alternate way is to modify the scoring functions such that $t + F$ is replaced simply by t. This will impact the quality of the results. However, we believe that an intelligent post processing step will be able to mitigate the loss in accuracy. This step will have an flavor of online approximation algorithms. We are curretly looking into this aspect.

Actor Names	Career Tenure
Tom Kenny	1989-2008
Ron Perlman	1975-2008
Leonald Martin	1989-2008
Samuel Jackson	1987-2008
Greg Ellis(I)	1997-2008

Table 19: Top Actor Births in the IMDB dataset

Actor Names	Year of Split
Jerry Nelson, Dave Goelson	1985
Douglas Newell, Lillian Carlson	1995
Lyndon Johnson, Martin Luther King	2000

Table 20: Top Actor Splits in the IMDB dataset

Actor Names	Year of Merging
Hulk Hogan, Kevin Nash	1996
Corey Burton, Frank Welker	1998
Rossie O' Donell, Jay Leno	1998
Ron Jeremy, Alex Sanders	2000
Sid Caesar, Howard Morris	1998

Table 21: Top Actor Merges in the IMDB dataset

2. **Study the changes in transformation score over time:** This study can be extremly interesting and provide more details about the evolution and interaction among the features. For example, Figure 2 shows how strength of various actors vary over time. *Bush* and *Kerry* are clearly the strongest actors in this corpus. An extremely high peak was observed for a small time period for *Dean*, but it ceased soon. *Abortion* shows intermittent peaks, corresponding to debates/speeches delivered by candidates. *Vietnam* was a strong actor for a small time period (with Kerry replying to Vietnam related remarks).

3. **Using domain knowledge to improve scoring metrics** : The current scoring mechanism does not incorporate any additional domain based inputs. For example, in the DBLP data, information on the impact of a publication can be used to improve parameters like Strength and Size of a cluster of authors. The impact can be measured using standard productivity metrics based on conference Ranking and citation Analysis [1]. Similarly, in the IMDB data, the movie ratings information can be used to calculate the impact of a movie, and used to fine tune some of the parameters using for scoring.

4. **Temporal Summarization of Evolving Datasets** : There has not been much work, relating to the topic of temporal summarization of time varying datasets. In our earlier work we showed how the ranked transformations can be used to summarize news graphs. The summarization is similar in spirit to Allen's temporal algebra [2]. We are currently generalizing the ideas.

5. **Visualization of Evolving Data** : While there has been much work in the field of visualization of static graphs, here has been very little in the field of visualizing dynamic graphs, and that too, only in the topics relating to better layout and energy minimization techniques [7]. Mining important events in evolving data, presents a great way of visualizing critical events, while varying the parameter of criticality, to view less important ones.

6 conclusion

In this paper, we presented a general scheme for quantifying changes in time varying data. Specifically, we focussed on changes derived from evolution and interaction of features inside such data. We described five key transformations, *create, cease, merge, split and continue*, in order to characterize the changes. Similar transformations have been proposed in seperate research works in different temporally evolving domains, like social networks, text and bioinformatics. We proceeded to define a scoring mechanism in order to find the most important transformations out of the discovered set. The motivation derived from the observation that the number of discovered transformations can be large for a huge dataset, and hence a user would have to manually sift through them in order to find the interesting ones. We based our scoring mechanism on three general parameters, that we believe,

Actor Names	Tenure of working together
Frank Welker, Michael Bell	1987-2004
Nick Nicholson, Henri Strzalkowski	1979-2004
Jim Henson, Frank Oz	1970-1990
Ron Jeremy, Tom Byron	1983-2005
Shawn Michaels, Mark Yeato	1988-2008

Table 22: Top Actor Continues in the IMDB dataset

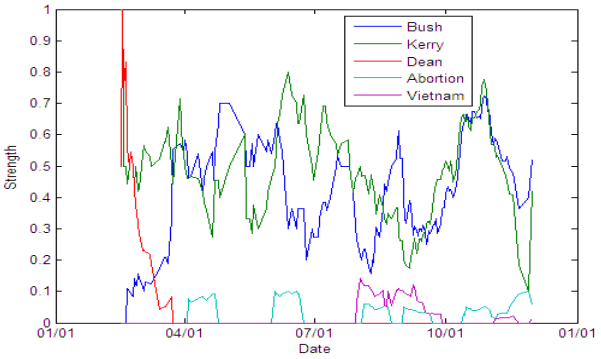


Figure 2: Strength over time for various actors in US Election 2004

characterize the spatio temporal properties of a feature vis-a-vis the corpus and also internally. The parameters were combined together to form scoring metrics for the five transformations, with the freedom of choosing combination weights based on the domain. We described our experiments with four different temporal datasets from the domain of text, social networks and scientific data. We empirically showed that the top tranformations do correspond to interesting and informative events in the data, and also presented a timing analysis of the algorithm.

References

[1] The Thomson Corporation. How do we identify highly cited researchers?, as of 2003.

[2] James F. Allen. An interval-based representation of temporal knowledge. In *IJCAI*, pages 221–226, 1981.

[3] The internet movie database: http://www.imdb.com/interfaces.

[4] Sitaram Asur, Srinivasan Parthasarathy, and Duygu Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 913–921, New York, NY, USA, 2007. ACM.

[5] Sitaram Asur, Srinivasan Parthasarathy, and Duygu Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. In *KDD*, pages 913–921, 2007.

[6] Rohan Choudhary, Sameep Mehta, Amitabha Bagchi, and Rahul Balakrishnan. Towards characterization of actor evolution and interactions in news corpora. In *ECIR*, pages 422–429, 2008.

[7] J Ellson, E Gansner, E Koutsofios, S North, and G Woodhull. Graphviz and dynagraph - static and dynamic graph drawing tools. In *Graph Drawing Software*, pages 127–148, 2003.

[8] Ming Jiang, Raghu Machiraju, and David Thompson. Geometric verification of features in flow fields. In *IEEE Visualization*, 2002.

[9] George Karypis and Vipin Kumar. *MeTis: Unstrctured Graph Partitioning and Sparse Matrix Ordering System, Version 2.0*, 1995.

[10] Gautam Kumar and Michael Garland. Visual exploration of complex time-varying graphs. *IEEE Trans. Vis. Comput. Graph.*, 12(5):805–812, 2006.

[11] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187, New York, NY, USA, 2005. ACM.

[12] Sameep Mehta, Steve Barr, Tat-Sang Choy, Hui Yang, Srinivasan Parthasarathy, Raghu Machiraju, and John Wilkins. Dynamic classification of defect structures in molecular dynamics simulation data. In *SDM*, 2005.

[13] Qiaozhu Mei and ChengXiang Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD*, pages 198–207, 2005.

[14] Qiaozhu Mei and ChengXiang Zhai. A mixture model for contextual text mining. In *KDD*, pages 649–655, 2006.

[15] Joshua OMadadhain and Padhraic Smyth. Eventrank: A framework for ranking timevarying networks. In *LinkKDD*, 2005.

[16] Deborah Silver and Xin Wang. Volume tracking. In *IEEE Visualization*, pages 157–164, 1996.

[17] Myra Spiliopoulou, Irene Ntoutsi, Yannis Theodoridis, and Rene Schult. Monic: modeling and monitoring cluster transitions. In *KDD*, pages 706–711, 2006.

[18] Hui Yang, Srinivasan Parthasarathy, and Sameep Mehta. A generalized framework for mining spatio-temporal patterns in scientific data. In *KDD*, pages 716–721, 2005.

[19] Hui Yang, Srinivasan Parthasarathy, and Sameep Mehta. Mining spatial object associations for scientific data. In *IJCAI*, pages 902–907, 2005.

[20] X. Yang, S. Asur, S. Parthasarathy, and S. Mehta. A visual-analytic toolkit for dynamic interaction graphs. In *To appear in KDD*, 2008.