III-CXT: Collaborative Research: Computational Methods for Understanding Social Interactions in Animal Populations

1 Introduction

Which individual will animals follow when moving away from a predator? When one of the animals leaves the population, will it affect the entire social structure? Which females are likely to form a harem? Will a group of animals move together or disperse when their territory is destroyed? For animals that live in groups, social interactions and structure play a key role in their response to changes. Yet, the impact of this structure on the behavior, ecology, and evolution of animals has not been addressed in population biology. One of the biggest obstacles to studying social structure in animal populations has been the lack of technology and methodology to collect and analyze the necessary data.

The goal of this research is to develop computational tools that will allow biologist to exploit the imminent, widespread availability of sensor-derived data in order to analyze social structure of animal populations and extract patterns of social response to external perturbation events.

Nowhere is the impact of social structure likely to be greater than when species come in contact with, or attempt to avoid, predators. Prey have evolved traits to avoid detection and capture. These traits include physical adaptations, such as increased speed and improved eyesight, and social adaptations associated with living in groups where coordinated action enhances hunting by predators and evasion by prey. Details about how physical adaptations function abound [96]. But little is known about how the structure of prey societies affects the likelihood of prey avoiding detection or capture. We can envision common situations in which the success or failure of a predation event is determined by the history of interactions among individuals within prey groups. For instance, if each member of a prey group has interacted recently with most other members, then an alarm signal by one who has sighted a predator is likely to spread quickly through the group. By contrast, when individuals in groups are bonded weakly, such coordinated behaviors are less likely to develop. If a prey group comprises several tight cliques that have only recently come together, then individuals may respond only to the vigilance or flight of those within their clique. Thus the structure of an animal species' society should affect its ability to avoid becoming prey.

Until very recently, addressing such biological concerns would have been impossible. Populations contain intricate connections that change on time scales ranging from minutes to generations. Yet, traditional field data in ecology and population biology studies came from direct visual observation, using the standardized approaches of scan or focal sampling [4, 84]. In scan sampling, observers record a series of "instantaneous" snapshots of location of each individual in a group. This method provides simultaneous estimates of position and activity for many individuals but precludes fine-scale continuous recordings on many individuals and thus misses rare and brief yet important events. In focal sampling, an observer records all behaviors of interest for one or few individuals at a time, thus trading off high resolution for replicated synchrony. To demonstrate the need to overcome the limitations of current sampling methods, consider a population with infrequent predation or an elusive predator. Each time there is a predator the population forms a distinct social pattern, with males, lactating females and juveniles in certain positions. The individual animals may be in different position each time but the overall social structure is the same. Scan observations might miss the infrequent predation events altogether, while focal observations focused on a particular individual will not discern the global social structure. For a subtler example, consider a hypothetical population made up of groups that frequently exchange one individual member at a time with different other groups. The turnover in these groups is actually very low yet after many exchanges

the groups have completely different membership. Moreover, if the exchanges are random then the groups will ultimately have random membership. Infrequent scan observations will conclude that the structure of the population is completely random. Focal sampling of the associates of one individual will produce a picture of a reasonably stable group followed by a jump to another stable group. Neither describes what is really going on in the population.

In both examples, in order to identify patterns of social interactions and answer questions regarding relationships between social contacts and environment, biologists need the means to collect and analyze continuous, fine-scale data on multiple individuals simultaneously. Biologists are beginning to address data collection limitations by deploying sensing devices on animals to automatically collect data via wireless networks. These sensors are embedded within, or are attached to, individuals and include Global Positioning System (GPS), heart-rate monitors, video cameras, and audio recorders [78].

For social contact data in many fields of study, researchers typically employ a network framework, in which individuals are linked if they have interacted [81]. To estimate strength of pairwise interaction, researchers may aggregate data over months or years [32]. Again, data aggregation presents a problem for testing hypotheses about how network structure reflects responses to singular short events, such as predation. For example, a network created by aggregating a week of data will give little insight into the social response to a predation event that happened in the course of half an hour. Time scale mismatch may thus result in inaccurate inferences. Lacking concepts to analyze the dynamic relationship between social contacts and external events, biologists are both unable to make full use of temporal information in existing data and unprepared for the imminent, widespread availability of sensor-derived data. In this proposal we present a conceptual framework and algorithmic tools for meeting these challenges.

2 Research Objectives

We propose novel conceptual and algorithmic solutions for analyzing sensor-derived data on social contacts and predicting patterns of social interactions. Our methods will incorporate information about the timing of contacts of multiple individuals and work within the resource constraints of the sensor-based data collection method. We will then test the accuracy and predictive power of our algorithms by characterizing the social structure of horses and zebras (equids) both before and after human- or predator-induced perturbations to the social network. Our goal is to design computational techniques to identify entities such as a community, leaders, and followers. We will predict social response patterns to danger or disturbances, and demographic changes in the population. We will focus on rare but significant events and critical individuals. Our approach is to combine ideas from social network analysis, Internet computing, distributed computing, graph algorithm design, and machine learning to solve problems in population biology, both animal and human (e.g., epidemiology). The ultimate goal of our interdisciplinary research is to design a powerful and general abstract mathematical model that captures the distinct properties of dynamic interaction networks and rigorous, efficient, and scalable algorithms to answer queries within this model.

We will evaluate our approach using a case study on equid species-both domestic and wild– that have complex, dynamic social contact patterns and move over large spatial scales. As a group, horses are especially useful because they include domestic animals, which are tractable for experiments. Wild species, on the other hand, live in varied habitats across the globe and provide important natural variation in social structure. There are several advantages to using animals: spatial proximity in general sets the stage for most social interactions. Thus, we will use GPS location to derive spatial proximity information, which we will use as the first approximation of

social interactions. In addition, we can simulate some of the conditions corresponding to the events of interest, such as predation, to verify the accuracy of our computational techniques. While we will work with equids, the approach can be broadly applied in any species, including humans, to questions about patterns of social interactions and their relationships to external events. The tools we will develop are appropriate for both the anticipated large datasets from new sensor networks as well as for existing data on interactions and activity at coarser temporal resolution. We organize our computational research goals into two components: 1) dynamic social interaction analysis to extract qualitative information about communities and specific individual roles and 2) developing predictive models of social interactions and using these models to identify social response to external perturbation events.

3 Preliminary Work and Results

We now describe the work we have already accomplished and the results we have obtained as a proof of concept of our approach.

1. Populations consisting of hundreds of plains and Grevy's zebras were observed over a period of 7 years (1999-2006) on a number of different commercial ranches and conservancies in the Laikipia region of Kenya. At each location predetermined census loops were driven on a regular basis (approximately twice per week) and individuals were identified by unique stripe patterns. Upon sighting, an individual's GPS location, behavior, the identity of associates and a variety of habitat features characterizing ecological context were recorded and entered into a database. Detailed behavioral measures were recorded using either focal or scan sampling methods and were combined with ecological measures to determine the causes and consequences of herding. Movement routes and home ranges were estimated from linking together repeated GPS locations or from more frequent GPS fixes derived from our Zebranet GPS tracking collars.

2. The data have been analyzed using the currently standard aggregate static analysis tools [87]. We have been able to show that herds form mostly to reduce the chance of cuckolding by marauding bachelor males [81]. Based on repeated observations of known individuals, we know that lactating females are responsible for initiating movements that lead to water [32]. By combining visual observations with movement data gathered by Zebranet GPS tracking collars, we have also been able to demonstrate that zebras adjust the speed of movement and magnitude of turning by time of day and habitat features to reduce the risk of predation [33].

   Association data derived from repeated observations of individually identified zebras have generated static social networks for broadly similar fission-fusion species such as Grevys zebras and wild asses. In these species individuals change associates frequently and Figure 3 of this proposal illustrates one such network for Grevy's zebra. When the static networks are compared [87, 82] subtle differences in structure are revealed. Whereas Grevy's zebras show higher degrees of cliquishness and females of similar reproductive state are more likely to stick together then in wild asses. Overall, network analyses suggest that Grevy's zebras are more selective with respect to whom they associate which is likely the result of the greater risk of predation they experience. While network theory has provided essential tools for characterizing quantitatively social structure, the averaging of data over long periods most likely obscures important relationships that are more aligned with processes of information and idea transfer that are responsible for holding societies together.

3. We have developed the initial conceptual framework for analysis of dynamic social networks [9] (Section 4.2 describes the details of this framework). We have used the framework to

identify dynamic communities in Grevy's and plains zebra populations from the initial data described above. The identified communities correspond to harems, supporting our intuition. However, the datasets are too small, the time resolution is too coarse, and every timestep has many missing individuals. Moreover, there is no information on the external events that may affect the population. Thus, it is currently not possible to use the datasets for deeper analysis. Hence, we need to collect data systematically using GPS collars on a set of individuals at a fine temporal resolution, with all the individuals being observed at all timesteps and the correspondence to the population perturbation events being retained.

Within the conceptual framework, currently only the basic algorithms are implemented. While we have a theoretical understanding of the algorithmic solutions to some other problems of interest, such as finding critical individuals, interactions, and timesteps (for some context of "critical"), these algorithms have not been implemented or tested against a real data. Moreover, some of these algorithms do not scale up to large datasets even in theory. Furthermore, there are still many open questions that we have not yet addressed conceptually, such as identifying a leader or characterizing changes in social structure.

4. We have developed the initial algorithm for predicting structure of social interaction patterns [58] (see Section 4.3 for the description of the method). We have used the initial plains zebra data to assess the accuracy of the approach. The algorithm predicted the pattern of interaction with over 80% accuracy. Again, we could not use the Grevy's data since it is too small and incomplete. The algorithm will serve as a solid foundation for developing data mining tools for extracting sequences of interaction patterns characteristic of population perturbation events.

4 Proposed Work

The ultimate goals of this interdisciplinary research are: to design a powerful and general abstract mathematical model that captures the distinct properties of dynamic social networks; to design rigorous, efficient, and scalable algorithms to answer queries within this model; and to validate the model and our algorithms using domestic and wild equid populations. Our computational research goals are twofold. First, we will design algorithmic techniques to identify social entities such as a communities, leaders, and followers. We refer to this set of tools collectively as "Dynamic Social Interaction Analysis" and describe our approach in Section 4.2. Second, we will design algorithms to predict social response patterns to danger or disturbances, as described in Section 4.3. We anticipate significant overlap in the algorithmic tools that will be useful in achieving these two goals. Below we outline the details of each of the components.

4.1 Data collection and validation using zebra populations

In order to test the efficacy of the algorithms developed for characterizing and predicting dynamic social interactions, we will carry out two sets of field studies. The first will entail observing the behavior of domestic horses in barnyards, recording proximity and social behavior among known individuals over time. Proximity measurements will provide the data for constructing networks and forecasting network dynamics, whereas data from social interactions will be used to identify critical individuals. Characterizing individual horses as dominants or subordinates, leaders or followers and cooperators or competitors will provide three different social dimensions in which critical roles may emerge. Horses boarding in stables provide a unique means of measuring the impact that critical individuals can have on social structure. The herds that form in barnyards often experience short- and long-term social perturbations as individuals are taken for rides or transferred among stables. By working with owners and stable managers we will be able to remove specific individuals that

our analyses identify as "critical" and measure the impact of their removal on network structure. In this way the first part of our fieldwork will be used to "ground-truth" the accuracy and power of our dynamic analyses.

The second feature of our fieldwork will involve studying how the dynamics of plains zebra herds respond to the presence of predators, a natural perturbing force. Although wild horses and plains zebras live in core social groups whose membership remains fairly constant, groups of domestic horses in barnyards and natural herds of zebras consist of agents whose associations change regularly. For domestic horses the agents with fluid associations are individuals, whereas in plains zebras they are the harem groups themselves. Previous research [81] has shown that herd size and composition are affected by degree of sexual harassment imposed by bachelor males and the abundance of food and predators. In this part of our study we will place GPS tracking collars on one member of many zebra harems and lion prides in a nature conservancy in Kenya. In this way, we will remotely monitor the movements and associations of zebras as well as their predators over time as they move among different habitats during day and night. Our past research shows that plains zebras change their habitat preferences and move more erratically – higher speed, sharper turns – at times of peak lion activity [33], but we do not know if they change their associates or types of associates as well. Ultimately, our dynamic network analysis tools will be able to answer this question. Since we expect social structure to evolve to afford prey protection against predators, we will expect to see changes in associations and network structure as the risk of predation changes with time of day (and habitat location). Thus, we will be able to use this time-correlated data to test the accuracy of our predictive algorithms.

We will gather behavioral data on a subset of zebra harems similar to that gathered on domestic horses. With these data, we will assess whether "critical" harems exist, and whether there are "critical" individuals within harems. We know that certain individuals take on leadership roles in the context of group movements. In the tight-knit groups of plains zebra harems, for example, particular females emerge as consistent harem leaders, while in loose-bonded herds there are no such long-term leaders among harems [32]. We will use this difference to test the reliability of our algorithms for predicting leaders and critical individuals.

Thus the history of interactions among individuals or subgroups influences the development of habitual roles. If we find critical individuals or harems in the association network, we can then assess whether these individuals or subgroups organize herds when responding to predators or when reducing predation risk by preemptively adjusting behavior as ecological circumstances change.

4.2 Dynamic Social Interaction Analysis

Finding patterns of social interaction within a population has been addressed in a wide range applications including: disease modeling [29, 28, 43, 56, 66, 65, 77], the web and other information networks [5, 27, 30, 50, 54, 57, 72, 75, 86], cultural and information transmission [8, 17, 19, 22, 45, 89, 90, 94], intelligence and surveillance [1, 8, 55, 63, 62], business management [12, 21, 73, 74], scientific collaborations [7, 15, 14, 60, 68, 67, 76], conservation biology and behavioral ecology [23, 26, 61]. Until recently, questions regarding social structures and social interactions in human or animal populations were answered using statistical tools which summarized aggregate information over the population over long periods of time.

Traditionally, the main measure of social structures in animal populations has been an index of association for all pairs of individuals based on the proportion of time spent in the same group [18, 35, 95]. Thus the entire time series of data is compressed into one number for each pair. Typically, biologists use statistical tools to measure the significance of the association patterns and their correlations to demographic groups and the environment.

Recently, social networks and traditional social analysis techniques are starting to be introduced

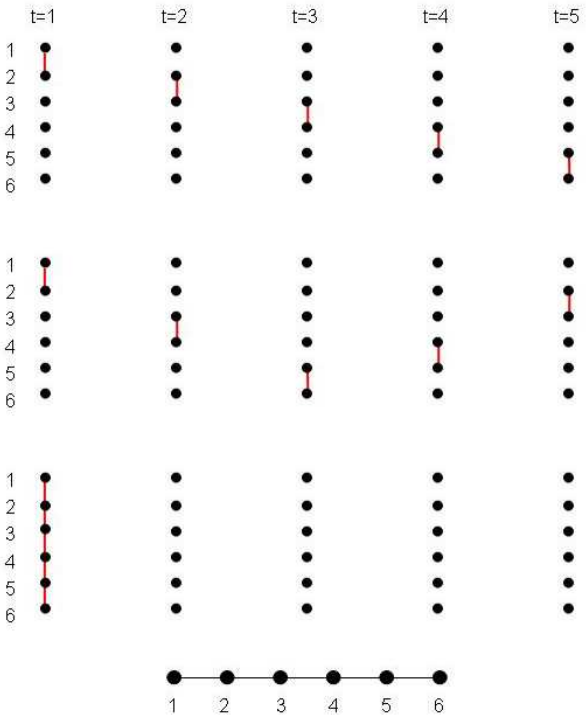

Figure 1: The top three blocks represent social interaction scenarios over five timesteps. The bottom row represents the same static graph obtained for all three dynamic scenarios. In the information dissemination setting, the graph represents potential transmission edges. Suppose we are interested in finding the most influential individual so as to maximize the total number of individuals possessing information at the end, assuming any contact will facilitate the transmission with some probability. In the static graph there is no directionality to the edges and the transmission can proceed both ways so a central individual (3 or 4) is most likely to maximize the spread. However, in the first dynamic scenario information cannot be transmitted from higher numbered individuals to the lower numbered ones (except once to its immediate neighbor). Thus the individual that will spread the information to largest number of others is individual 2. Similarly, individual 4 is the highest spreader in the second dynamic scenario. In the third scenario the individuals come in contact with each other only once. Thus, any individual with two neighbors will pass the information to at most two other individuals, which is the highest possible in this case.

to population biology. Cross et al. [26] show, using heuristic simulations on a static social network of American buffalo, that it is necessary to take the timing of social events into consideration. Lusseau and Newman [61] use a static social network to study the community structure in dolphins. Darren Croft and colleagues study the static social network of guppies [23, 25, 24] and we have done the same for zebras and wild asses [87, 82]. However, none of these approaches addresses the dynamic aspects of social interaction.

Social networks have been used as a sociological model of human interactions for several decades [34, 36, 40, 51, 52, 53, 69, 70, 71, 91, 92]. The recent increase in the amount and detail of data on social interactions (especially through electronic communications) has necessitated development of systematic computational approaches to social network analysis [97]. The network model of social interactions has been successfully used in many applications. However, a major drawback of this model is that it is essentially static in that all information about the time that social interactions take place is discarded. The static nature of the model gives rise to two major problems.

First, it can give us inaccurate or inexact information about patterns in the data, as Figure 1 illustrates. This example shows that determining critical individuals based solely on static data can give incorrect results. A second, perhaps more serious, problem with the static graph representation is that it prevents us from even asking certain fundamental questions about either the causes or consequences of social patterns. What types of social interactions occur during a predation event? How quickly can information or a disease spread through the population? How do the size and stability of social structures change with outside circumstances (e.g. season, time of day, predator activity, upcoming conference or journal deadlines, court subpoenas, terrorist activities)? Are there differences in the life span of social structures with respect to their size and the demographics of their members? To be able to answer these questions, we need to have information on when social interactions occurred, specifically with respect to external events.

The emerging approach to dynamic network analysis has been proceeding in several directions. The statistical mechanics view [3, 6] considers networks as complex physical systems and strives to describe laws governing their evolution and limit behavior and properties. A more computational view [20] incorporates probabilities and uncertainty into the structure information and combines social network analysis with multi-agent systems. Computer simulations until recently have been the main computational technique to incorporate dynamic network information, e.g. [28]. Last few years have seen a development of systematic algorithmic approaches to dynamic network analysis, mostly in the context of information networks [2, 45, 52, 50, 57, 59]. Yet, most of the methods focus on the *frequency*, rather than *concurrency* and *order* of interactions.

To address the dynamic aspect of social interactions we use a network population model to capture time snapshot information. We then build a time series graph that connects the snapshot information. The connections within a time snapshot are the individual associations that exist at that moment. We impose a metric on sets of individuals within different time steps that allows us to track changes in social structures over time. In this model [9], many questions about the social population dynamics become classical questions of graph connectivity. Some of these are easily answered, yet many pose challenging algorithmic problems. We now describe this approach and our assumptions in more detail.

We assume that at any given instant the population is partitioned into groups of interacting individuals and that any individual can belong to only one such group. Thus, we develop our conceptual framework under the assumption that the input is in the form of the partition of the individuals into groups at every time step. This may be a result of the direct observation data or the output of processing of the pairwise connection data.

Given a population $X = \{x_1, ..., x_n\}$, we define a *group* to be a subset $g \subseteq X$. We assume that the input is a set of partitions (time snapshots), $P_1, P_2, ..., P_T$ of X, and that each partition, P_i, is a set of disjoint *groups*.

$$\forall i, \quad 1 \leq i \leq t \quad P_i = \{g_{i1}, ..., g_{ii_m}\}, \quad \cup_{j=1}^{i_m} g_{ij} = X \text{ and}$$

$$\forall p \neq q, \quad 1 \leq p, q \leq i_m \quad g_{ip} \cap g_{iq} = \emptyset$$

We denote by $P(g)$ the index of the partition to which g belongs. That is, if $g \in P_i$ then $P(g) = i$.

Given two groups, g and h, a set similarity measure $sim(,)$, and a *turnover threshold* β, the two groups are *similar* if $sim(g, h) \geq \beta$. Our definition here is independent of the set similarity measure. However, for the purposes of the study we will use an extension of the standard Jaccard similarity measure [41], namely $sim(g, h) = \frac{2|g \cap h|}{|g|+|h|}$.

We now define the main concept of our framework - a metagroup.

Definition 1 *Given an input in the form of partitions $P_1, ..., P_T$ of a set of individuals X, a set similarity measure $sim(,)$, a turnover threshold β and a function $\alpha(T)$, a* metagroup *MG is a sequence of groups $MG = \langle g_1, ..., g_l \rangle$, $\alpha(T) \leq l \leq T$ such that*

1. *no two groups in MG are in the same partition and the groups are ordered by the partition time steps: $\forall i, j, \quad 1 \leq i < j \leq l, \quad P(g_i) < P(g_j)$,*
2. *the consecutive groups in MG are similar: $\forall i, \quad 1 \leq i < l, \quad sim(g_i, g_{i+1}) \geq \beta$.*

We call the parameter α the persistence *of a metagroup.*

In practice, the questions defined below would be address over a range of values for α and β to discover significant metagroups.

Note, that the intersection between g_1 and g_l may be null by this definition; our only constraint is that the groups change gradually (as defined by β).

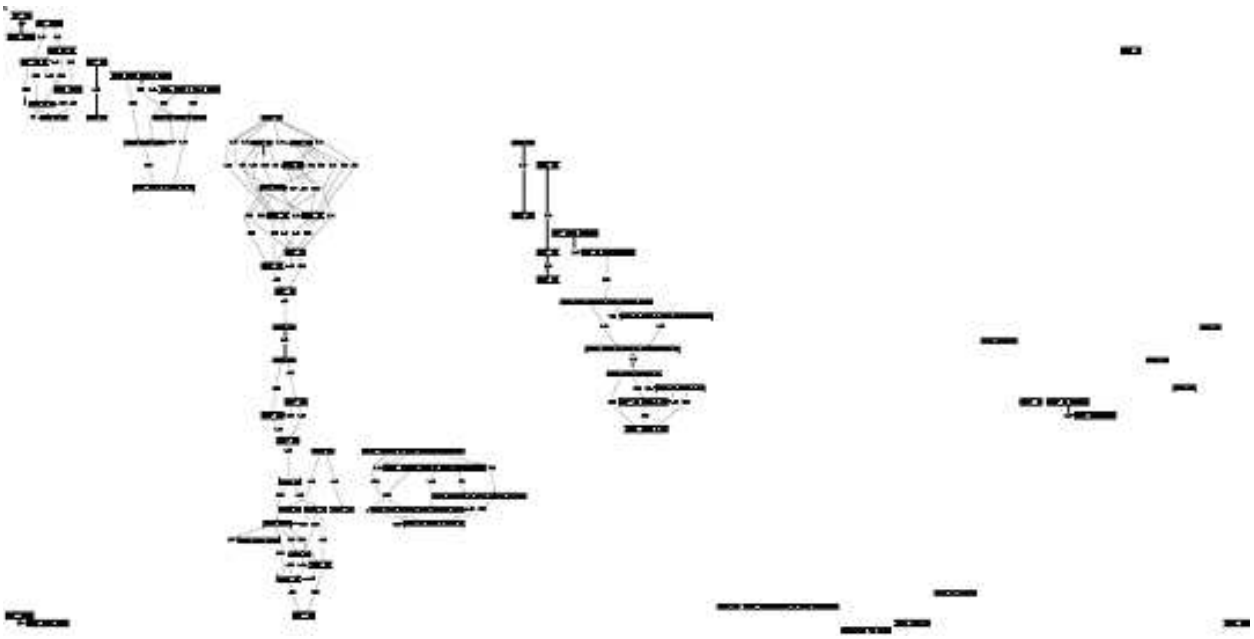

Figure 2: Metagroups graph of the Grevy's zebra population. The timeline moves from top to bottom. The vertices (represented by bars) at each horizontal level correspond to groups of individuals observed at that timestep. Edges connect groups with at least 60% similarity (β=.6) that are no more than 3 timesteps apart.

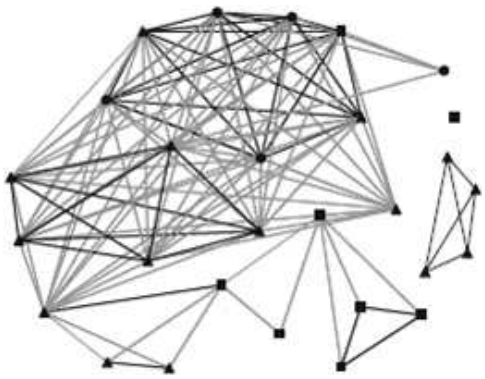

Figure 3: Observed static network for Grevy zebra (28 individuals). Individuals are vertices, with reproductive status indicated by shape: males (squares), lactating females (circles), and nonlactating females (triangles). Thin gray lines join individuals observed together at least once (nonzero network). Thick black lines represent statistically significant associations (preferred network).

Definition 2 *An individual $x \in X$ is a* member of a metagroup $MG = \langle g_1, ..., g_l \rangle$ *if the number of groups $g_1, ..., g_l$ to which x belongs is at least γ; where γ is some a priori chosen* membership threshold *function γ (which may be constant or a function of T, the total number of individuals associated with MG, and other parameters).*

In practice, the range of values of γ defines the subsets of individuals in a metagroup from the core to the periphery.

We use a weighted multipartite directed graph for the conceptual representation: $G = (V_1, ..., V_T, E)$ where V_i is the set of groups in partition P_i and $(g_i, g_j) \in E$ if $P(g_i) < P(g_j)$ and $sim(g_i, g_j) \geq \beta$. Note that this is a directed acyclic graph (DAG) since all the edges are directed from an earlier time step to a later one. The weight $w(g_i, g_j) = sim(g_i, g_j)$. A metagroup in this graph is a path of length at least $\alpha(T)$. We shall call this graph a *metagroup β-graph.*

Figure 2 shows an example of a metagroup graph for the Grevy's zebra population and Figure 3 shows its traditional static representation. In the static network, most of the individuals form

a large interconnected component with some peripheral single individuals or small groups. Even the statistically significant associations show two large communities and many fringe individuals. However, the metagroup graph reveals that there are five different communities. Moreover, while all of these persist for some time, the duration (α) of these communities ranges from 5 to 22 timesteps.

4.2.1 Metagroup Statistics

We can compute statistics describing a dynamic network that correspond to biological questions. We provide examples here of metrics that will be of particular interest for our study subjects.

While it might be impractical to list all the metagroups in some cases (since there may be an exponential number), there are nonetheless many statistics we can calculate efficiently. For example, **the number of metagroups** is the number of paths of length at least α in the graph and the **average metagroup length** is the average length of a path which is at least α long. Both can be computed with a simple dynamic programming algorithm.

The **most persistent metagroup** is the metagroup that exists for the longest number of time steps. Finding the most persistent metagroup is equivalent to finding the longest path in a DAG, which is a well-studied problem that can be solved in linear time using dynamic programming on a topologically sorted graph. Finding this group should be helpful in answering questions about persistence of social connections in different animal populations; determining persistence of social connections is a fundamental problem in population biology. Using this metric we will have an unprecedented opportunity to test hypotheses about how the types of relationships that develop in a society depend upon persistence of social interactions. For example, is it only those individuals involved in long metagroups that form cooperative relationships as exhibited by grooming behavior?

Although the algorithms for computing these fundamental statistics about the dynamic properties of social connection are simple, they nonetheless are unacceptably slow on large data sets. The challenge is to design sublinear algorithms that do not look at the entire input. While the statistics about the entire population over the entire period of time do indeed require examining the entire input, questions about a subset of individuals or timesteps can and should be answered significantly faster. To do so, we will use the information we already have about the whole population and will focus on the intersection defined by the subset of individuals or groups under question. In many cases, such approach is faster than computing subset statistics from scratch each time.

4.2.2 Critical Groups

Unfortunately, not all questions about dynamic social networks can be solved efficiently using known algorithmic techniques. For example, the problem of determining the fragility of a social network can be formulated in many ways. One way is to find the smallest set of groups whose absence would leave no metagroups. We formalize this problem as the Min k-Path Vertex Shattering Set problem: for an arbitrary graph $G = (V, E)$ find the smallest (weighted) subset of vertices $U \subseteq V$ such that the subgraph induced by $V - U$ has no paths longer than $k - 1$. We have shown that the complexity of this problem very much depends on the specific value of k [9]. When $k = 2$ the problem is equivalent to Max Independent Set, which is a well-studied NP-complete problem. On the other hand, it is solvable in polynomial time when $k = T$, where T is the length of the longest path in the original graph. This dichotomy presents a theoretical challenge of analyzing the complexity for the entire spectrum of values of k. In this project, we will investigate several possible heuristics for finding critical groups and will compare various notions of criticality.

4.2.3 Critical Individuals

Another way to address the question of the fragility of a social network is to ask what is the smallest number of *individuals* whose absence would leave no recognizable social structure. The

size of this set indicates population resilience to targeted or random loss of individuals. We can test whether these critical individuals are also leaders in other respects, such as in initiating directed movements by a group or otherwise changing group activity. Having identified critical individuals, we can determine if they have special phenotypic characteristics such as age or dominance. Using stochastic simulations we can find one such individual by removing each of the individuals in turn and determining which individual's removal most dramatically changes the social structure. However, this brute force algorithm is not practical when trying to find an optimal set of more than one critical individual. Moreover, a greedy approach also fails since the functions measuring the social structure (e.g. the number of metagroups) are not monotonic (this is in contrast to the function in [45] which does have monotonic properties since it is over a static network).

Possible formalizations of the notion of critical individual include:

- Individuals whose removal minimizes the number of metagroups containing any k groups in time step 1.
- Individuals whose removal leaves no metagroups or no large metagroups. (Note that in some extreme circumstances such a set does not exist)
- Individuals whose removal minimizes the number of disjoint metagroups

All of these are theoretically and computationally difficult problems. We will investigate their theoretical complexity as well as analyze the following heuristic approaches:

- Remove the individuals that appear in the intersections of the largest number of groups that are still connected by an edge
- Remove the individuals that remove the largest number or weight of edges
- Remove the individuals connected by the heaviest edges
- Remove the individuals that remove edges from the largest number of metagroups

We will compare our notions of critical individuals to the standard static notions of critical individuals based on vertex centrality by using simulations.

4.2.4 Critical Times

One of the more powerful demonstrations of our dynamic model is that we can pinpoint times when the social structure of the population changes significantly. Intuitively, when the patterns of interactions within population are stable the size of any edge cut in the metagroup graph remains relatively the same. When the interaction patterns change the number of edges may either go up (indicating more individuals connecting between different groups) or down (due to high turnover), depending on β threshold. Thus, we can detect the times of social upheaval by finding the difference in the edge cuts. We will develop efficient algorithms that track the size of edge cuts and will validate our ideas on real data by inducing a change in the social structure of our experimental population.

4.2.5 Future of dynamic network analysis in population biology

The dynamics of societies, populations and communities depends on the flow of ideas, disease, potential mates and maturing offspring. Many population processes are affected by the stability of, and the connections among, population subgroups: the conservation of rare and endangered species, the resistance of populations to the spread of virulent diseases, and the stability of communities as the abundance of species from different trophic levels change. Understanding what factors affect these flows and characterizing the nature of networks that facilitate or impede them will be critical for solving the problems described above.

4.3 Prediction of Social Interaction Patterns

A different approach to identifying patterns of social interactions is to find an explicit predictive model of the interaction patterns over time. The objective is, given a timeseries of observations of animal interactions, to predict exactly when certain interactions will occur in the future. If the predictions are accurate then the model used to make those predictions accurately describes the pattern of social interactions and can thus be used to identify significant patterns and sequences of patterns of interactions corresponding to population perturbation events.

We have designed a representation for this type of temporal data and a generic, adaptive algorithm to predict the pattern of interactions at any arbitrary point in the future [58]. We test our algorithm on predicting patterns in e-mail logs, correlations between stock closing prices, and social grouping in herds of Plains zebra. Our algorithm averages over 85% accuracy in predicting a set of interactions at any unseen timestep.

Temporal networks[1] are a powerful generic model for representing interactions over time amongst a set of individuals [44]. A temporal network consists of a sequence of regular graphs, each being a snapshot of interactions at a particular instant or over a small time interval. Vertices in each graph represent individuals and edges between them represent interactions, which can either be directed or undirected (bi-directional). The flexibility of the definition allows temporal networks to be used to model a variety of processes while maintaining the explicit order and concurrency of interactions.

There are many questions that can be posed for processes represented as temporal networks. We focus on the task of predicting the structure of the temporal network at *each timestep*, thereby computing a model of the evolution of the process under consideration. There is an extensive body of work that focuses on problems related to the evolution of networks, without the information on the order and concurrency of interactions. Such networks can be seen as collapsed aggregations of temporal networks. A closely related problem from such network analysis is that of *link prediction* in a graph, which aims to rank all possible edges (interactions) by the likelihood that they will occur in the future [60]. Our work extends that definition to temporal networks by predicting the structure of the graph which is not noise at each timestep. The predictions are based solely on prior observations. Note, that unlike the link prediction problem, we are concerned with the ability to predict exactly *when* groups of interactions will occur, not to predict the likelihood of every possible interaction occurring in the future, regardless of the timing and order of these occurrences.

We have designed a generic, accurate, adaptive, streaming algorithm for efficient structure prediction in temporal networks. Our algorithm does not rely on any domain-specific parameters. The data is assumed to be a stream of graphs, and the algorithm adaptively learns a model for the evolution of the process. Our algorithm uses the idea that probability density functions for the time interval between every pair of interactions can be used to make predictions about subsequent timesteps. However, directly using this approach requires computing on pairs of edges and becomes intractable for even medium-sized graphs. We propose the use of frequent subgraphs in order to aid the tractability of the algorithm as well as to filter out insignificant interactions. We tested our algorithm on three diverse examples of real-world processes, ranging from stock price correlations to dynamics in animal populations. Our algorithm averages 85.7% accuracy when predictions are allowed a single slack timestep. Figure 4 shows the accuracy of our algorithm on the Plains zebra dataset for different support values for the subgraph frequency. The plains zebra dataset consists of 1,002 unique individuals, 228 timesteps, and 1,415,815 total observed interactions.

Thus, we have designed an initial algorithm for accurately learning the pattern of significant

[1]Similar representations are also known as dynamic networks [16], time-series networks, and longitudinal data in social network analysis [85].

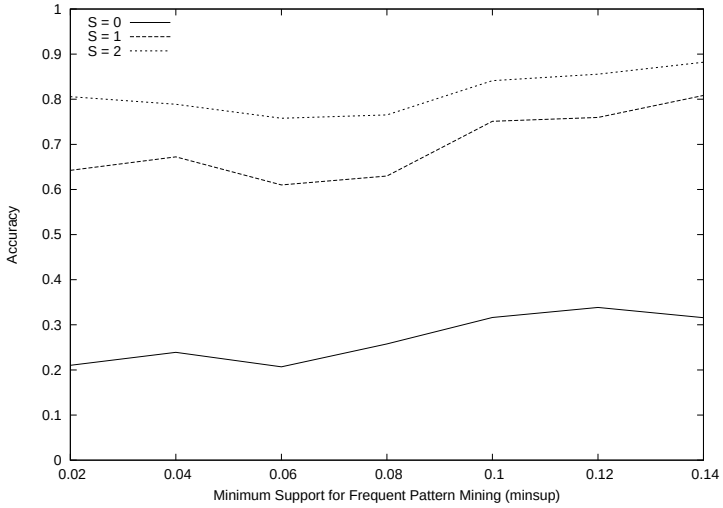

Figure 4: Prediction accuracy at different levels of minimums support for frequent subgraphs. The three curves correspond to different prediction timestep slack: at slack S a prediction made for timestep t is considered accurate if it occurs at timesteps $[t - S, t + S]$.

interaction over a period of time. This approach provides a good estimate of the underlying model of social interactions. Moreover, we can use this to test the accuracy of that model by measuring the accuracy of its predictions. Thus given sufficient data on predation response (and we will know when the data are sufficient by the accuracy of the predictions made using these data) we know what interactions are significant in response to those predation events. Moreover, we can compare those significant interactions to the significant interactions in absence of predation events to extract the pattern of social response to perturbation events, such as predation.

To identify sequences of interaction patterns, we will combine the generalized frequent subgraphs discovery with frequent sequence discovery [64] to find frequent temporal patterns of interactions. We will relate those dynamic interaction patterns to demographic and environmental factors using a data mining technique of rule association [39], which we will extend to accommodate time-series of attributes.

5 Broader Impact

We discuss both the research and the educational component of the broader impact of our proposal separately below.

5.1 Research

It is hard to overstate the applicability of our approach and its impact on many fields of study. Our conceptual tools will be useful for behavioral and social scientists from epidemiology to conservation biology, and disease ecology. The behavior classification methods we will develop are relevant to a wide range of studies in which researchers need to classify remotely sensed data into several categories of interest. Our network methods have broader relevance to human societies: disease transmission, dissemination of ideas, and social response to crises are all dynamic processes occurring via social networks. In the course of this project we will combine tools from several disciplines, including biology, sociology, computer science, and mathematics into a unified approach of studying mechanisms and implications of social behavior of a group of individuals.

5.2 Education and Training

Our research will provide the students in biology and computer science with hands-on experiences in asking and answering biological questions by developing new applications of computer science. The project will provide funding for interdisciplinary research by post-doctoral fellows and many

graduate students and undergraduate theses. It is our firm belief that undergraduate students should be involved and integrated into our research activities as much as possible. PI Berger-Wolf teaches a Computational Biology course, which is part of the required curriculum for Bioinformatics graduate majors at the University of Illinois, Chicago (UIC) and an elective course for computer science graduate students. She will incorporate research from this project into the course material. She will also use examples motivated by this research in her upper undergraduate and graduate level course Algorithm Design and Analysis. PI Berger-Wolf is actively involved in recruiting to and retaining women in computer science. From taking a team of 8 UIC students to the Grace Hopper Celebration of Women in Computing and founding the Women in Computer Science organization at University of Illinois at Urbana Champaign, to mentoring and numerous outreach activities, she has participated in every part of the process of bringing more women to computing. She has already given talks about computational approaches to population biology at a regional "Women in Computing" conference (aimed at girls from middle school level to undergraduates) and to the participants of the NSF funded Research Experience for Undergraduates at the Center for Discrete Mathematics and Theoretical Computer Science.

PI Berger-Wolf is committed to increasing interdisciplinary collaborations in science. She is organizing DIMACS workshop on "Computational Analysis of Dynamic Social Networks," which will bring together scientists studying the dynamic aspects of social networks in diverse contexts. The conference will facilitate exchange of ideas and the development of new approaches to explicitly address the time component of social interactions. The participants will include computer scientists, biologists, sociologists, epidemiologists, economists, physicists, information scientists, and mathematicians. The workshop includes specific support for graduate student and post-doc participation. The PI intends to make this workshop a regular event. The PI has given invited talks about the proposed study at the SIAM Conference on Discrete Mathematics, in Vancouver, BC, June 2006, American Mathematical Society Sectional Meeting's Special Session on Discrete Models in Biology, Johnson City, TN, October 2005, and at the Indiana University's Talk Series on Networks and Complex Systems. The PI will continue to promote interdisciplinary science by presenting this work at various cross-disciplinary forums. UIC is an urban institution with a higher percentage of Latino and African-American students than any other Big 10 university. It ranks 26th out of more than 2,000 colleges and universities in the number of Master's degrees awarded to Latinos and African-Americans, who together make up more than 23% of the student body. This allows an excellent opportunity to integrate diversity into research and educational activities.

Co-PI Rubenstein has had much success in bringing both women and minorities into his research projects. The co-PI teaches a number of undergraduate and graduate courses that cover topics in animal behavior. The aim of these courses is to demonstrate the power of approaches that meld theory and empirical research. Examples emerging from this study will become central to achieving this objective. They will serve as case studies as well as becoming important learning tools via problem sets and laboratory simulations. The co-PI is involved in many outreach efforts at Princeton. He serves on advisory committees of the Environmental Institute (PEI) and the Program in Teacher Preparation, both of which sponsor summer teaching institutes. One focuses on disadvantaged, and mostly minority, high school students from Trenton by bringing them to campus and immersing them in two summers of study to prepare them for college-level work. The program has not only improved the quality of colleges these students attend, but performance and retention is far above the national norm. Units on ecology, animal behavior and mathematics make up a large part of the program. The post-doctoral fellows and PI will use our results and models to teach some of these units. Discussing scientific and conservation issues related to charismatic megafauna always grabs the attention of students and makes a wonderful vehicle for showing how science works. Based on our work we will also be able to show students the power of mathematical

and theoretical analysis. The other summer institute focuses on teaching elementary school teachers how science works by engaging them in the "doing of science". We will use examples from our research to show how mathematical reasoning and computational analysis play key role in shaping scientific investigation, applied to real-world problems: collective behavior in large vertebrates.

Co-PI Saia is the chair of the recruitment and retention committee in the Computer Science department at the University of New Mexico. As such, he organizes several events each year to recruit high school and undergraduate students into Computer Science. Dr. Saia plans to incorporate research from this project into research demos to use during these recruitment events. These recruitment events will likely bring more minority students into computer science, as many of the targeted high schools are "majority minority" schools. Moreover, the University of New Mexico (UNM) is itself one of only two universities in the nation that is both an officially designated minority-serving institution and a Carnegie designated "very high research activity" institution. It is also one of the three largest producers of Hispanic and Native American engineers in the nation. In this project, Dr. Saia will strive to recruit minority undergraduate and graduate students to participate in research. The opportunity to apply computer science to study an endangered species will be a great attraction to such students.

6 Results From Prior NSF Support

Tanya Berger-Wolf

NSF IIS-0612044: Computational Methods for Kinship Reconstruction, 07/01/2006 – 06/31/2009. The goal of this research is to develop a robust computational method for reconstructing kinship relationships from microsatellite data. The work is ongoing. Thus far, the work has resulted in one submitted publication [11] describing a novel computational approach to sibship reconstruction in from same generation data in absence of parental information. A repository has been established for the project containing twelve microsatellite sample benchmark datasets and four methods for reconstructing sibling relationships. Three graduate students (one biology and two computer science) are being supported.

Daniel Rubenstein

NSF IBN-9874523: Multilevel social organization in plains zebra: From mating systems to social systems, 01/2000 – 12/2005 This study examines why herds of zebras exhibit such wide variation across landscapes, how the dynamics of herd formation are determined by the movement decisions of core social units and how the consequences of social relationships within and between breeding units determine individual reproductive success. Some major findings are [32, 33, 37, 38, 81, 80, 79, 80, 83, 87, 88, 93]: 1. Plains zebra herds mostly form to prevent bachelor males from harassing and mating with females. By banding together males amortize costs without lowering female foraging rate. 2. In addition, group composition is affected by preferences of females with young foals banding together to reduce the risk of predation. 3. Genetic analyses using microsatellite data show that herds consist of genetically related females but not males suggesting that females are the active players in constructing herds.

NSF IOB-0083827: Biocomplexity-Incubation Activity: Large Mammal movements through complex landscapes in East Africa, 10/1/2000 – 3/31/2004) This biocomplexity incubation study helped identify the problem and develop a research program for studying the problem of understanding the rules that govern animal movements and how human landscape changes will vary alter the porosity of the landscape and thus at times disrupt animal movements and their sustainability. Not only did it result in the submission of a full-fledged proposal to the biocomplexity program but it helped identify some of the technological needs and basic behavioral measures that form the basis of this proposal.

NSF DBI01-22373: FSML: Improvements in Facilities and Equipment at Mpala Research Centre, Kenya, 09/01/2001 – 08/31/2004 This project provided support for enhancing the research capabilities of the Centre. In a very real way they facilitated the research proposed in the current study. The following improvements were completed: a new laboratory that helped the DNA study, extra land rovers to help get to field sites, new computers and new solar electrification that insures power for experiments and data analysis. All have helped improve the amount and quality of research done at Mpala.

NSF CNS-0205214 ZebraNet: Position-aware Power-aware Wireless Computing for Wildlife Tracking, 09/2002 – 08/2007 This project has successfully developed a new generation of GPS tracking collars that are solar powered and use minimal electrical power to percolate data from collar to collar back to the base station [42, 98]. Two deployments have already shown that: 1. Zebra movements at night are different from during the day. Movements are more rapid and lead zebras to more forested habitats. 2. Zebra movements are affected by reproductive state. Stallions and females in harems move slowly over the landscape remaining in one square km area typically for 24 hours. Then they move quickly to water and to a new area where they repeat the pattern of intensive use. Bachelor males, however, patrol a core area everyday and make quick forays in search of females. That the collared bachelor male found the 3 collared females with stallions within 2 days shows the effectiveness of this strategy. 3. Zebras show strong habitat preferences that differ by time of day and by the reproductive state of the collared individual. Harmes use forests more at night then bachelors and they frequent watering points less often. All zebras prefer open areas with a few trees over open plains or bushland and forests.

Jared Saia

Professor Saia is the sole recipient of **NSF CCR-0313160: Scalable and Attack-Resistant Peer-to-peer Networks** which started in January 2003. In this grant, he has focused on 1) designing secure algorithms for distributed hash tables (DHTs) and 2) the preliminary stages of designing secure algorithms for computation in Peer-to-peer networks. This grant has had significant impact in both research and education. First, it has led to *seven* publications in some of the top conferences and journals in theoretical computer science [48, 49, 10, 31, 47, 46, 13]. The current number of citations to papers coauthored by Saia in this area is now over 300 according to both Google Scholar and Citeseer. Moreover, Professor Saia has been able to disseminate this research through other means. He has given invited talks at several workshops, universities and research labs. A result from his research on choosing a random peer has been incorporated into graduate classes at at least two universities. For example, Professor Aravind Srinivisan at the University of Maryland created a homework assignment and an exam question for a graduate class based on this result. Finally, Professor Saia has served twice, during this grant, on the program committee for *Principles of Distributed Computing (PODC)* which is the top conference in distributed computing.

Educational successes of the grant include the following. First, Vishal Sanwalani graduated in 2005 with a PhD under the advisement of Professor Saia. Vishal is currently completing a post-doc at the University of Waterloo, and is scheduled to begin a second post doc at Microsoft Research with Valerie King, a collaborator of Professor Saia. Professor Saia has also supervised five students who have completed Masters thesis including: Jake Proctor, now employed at Sandia Labs; Maxwell Young, now a PhD student at the University of Waterloo; I-Ching Borman, now employed at U. New Mexico Medical Center; Florina Cazacu, now employed at Waters Software; and John Alphonse. Florina and I-Ching are female and Vishal, I-Ching, Jake and Maxwell all graduated with honors. This grant has also helped support current PhD students Amitabh Trehan (expected to graduate in 2007) and Navin Rustagi (expected to graduate in 2008). In addition, Professor Saia has created a new graduate classes under this grant: "Algorithms in the Real World" and has also taught a graduate seminar entitled "Cybersecurity: A Theoretical Approach".

7 Coordination and Management Plan

7.1 Senior Personnel and Qualifications

We have formed a strong interdisciplinary team of researchers with a history of collaboration and cross-disciplinary research.

Berger-Wolf is an assistant professor of Computer Science, whose research has focused on computational population biology, combining her algorithms design background with experience in ecological and population modeling. Berger-Wolf will be the team leader, providing the connection between the biology and the computer science. She will formulate the biological questions and concepts as computational problems and together with Saia will work on designing the computational framework for analyzing dynamic social interaction patterns. She will work with the biologists (Rubenstein and Fischoff) on establishing the requirements for data and the methodology for validation of the computational tools on animal populations.

Rubenstein is a professor and chair of the Department of Ecology and Evolutionary Biology at Princeton, a world renowned behavioral ecologist who has been studying equids for over twenty years and has written definitive work on the subject. Rubenstein will be responsible for all biological aspects of the project, including biological hypothesis formulation, data collection and validation experiment design, as well as and directing the work of the postdoctoral researcher Ilya Fischhoff. He will also be in charge of the oversight and training of the undergraduate students and kenyan assistants in the gathering of the fine-grained behavioral data that will allow key individuals to be identified as "bosses", "leaders" or "cooperators", both as baseline and after perturbations.

Saia is an assistant professor of Computer Science whose expertise is in theoretical computer science with a focus on distributed algorithms and peer-to-peer networks. Saia's focus in this project will be twofold. First, he will focus on designing the computational framework for analysis of dynamic social interaction patterns. Second, he will focus on designing and testing rigorous, efficient and robust algorithms for answering queries in this framework.

Ilya Fischhoff will be the postdoctoral researcher in charge of carrying out the data collection and methodology validation experiments both in the US and Kenya. He will be a critical liaison between the data collection and the computational analysis and will work to get the appropriate data and to transform the raw data into usable form for the analyses.

7.2 Cross-Institutional and Cross-Disciplinary Coordination

We have formed a team of researches who are both comfortable working together and are strong independent researches in their respective areas of expertise. The extended Princeton University group will responsible for all biological aspects of the project. Berger-Wolf at the University of Illinois at Chicago (UIC) and Saia at the University of New Mexico (UNM) will focus on the computational aspects of project. Berger-Wolf and Rubenstein will be responsible for the overall intellectual coordination of the project.

To support this collaboration, we will have the following mechanisms (many of which are already in place):

1. We have and will continue to have bi-weekly (video)conferences using voice over Internet (VOIP) technology. The conferences will be both from the US sites as well as Kenya (depending upon the available bandwidth at the time of the conference). In Kenya, the research will be based at the Mpala Research Center. NSF (grant DBI01-22373) has played an important role in improving the site facilities and another improvement grant proposal is submitted. Together with other numerous benefits, this will improve the communication with the Center. A small operational and equipment cost is associated with the VOIP technology.

2. Berger-Wolf has a dedicated file server, that has dedicated disk space for the project to serve as the data, software, and publication repository. The space is accessible using a secure shell (SSH) protocol. All project members will have accounts on Berger-Wolf's lab computer system. The system has version control software installed to allow seamless editing and update by multiple users. The file server will be backed up (daily, retaining weekly and monthly copies) by the UIC Academic Computing and Communications Center (ACCC). We are *not* requesting funding to establish and maintain this repository.

3. We will have yearly meetings of all team members. Smaller groups of the team members will meet more frequently to work on various parts of the project. Postdoc Fischhoff will come to UIC for a week first and second year of the project to work on formulating the computational abstractions of the biological phenomena. Graduate students from UNM and UIC will travel to Princeton. All team members will meet at least once at each of the data collection sites. Berger-Wolf and Saia will meet at professional conferences. Travel funding is requested for these meetings.

4. A Masters graduate student (at 25%, written into the UIC budget) at UIC will work on technology transfer of computational techniques to biologists.

5. All universities have administrative staff familiar with inter-institutional interdisciplinary collaborative research project logistics. Staff support is budgeted into all submitted budgets.

7.3 Management Plan

Below we discuss specific timelines and goals, which will be reoptimized periodically for a fruitful completion of our project.

Year 1: Start collection of biological data on domestic horses. This requires training undergraduate assistants, constructing towers and video systems, gathering and analyzing data to determine animal "roles". Start collecting data from wild zebra populations in Kenya on animals facing natural predatory perturbations.

Continue collecting other available data sources (collaboration networks, cellphone data, blogspace etc.) and convert them to the format required for our study. Establish a usable data and software repository.

Develop theoretical framework for dynamic social networks, building on the existing work by the PIs. Develop and implement methods for identifying communities and critical times, individuals, and interactions. Develop initial measures of change in population interaction patterns. Develop and implement methods for identifying frequent characteristic sequences of interaction patterns. Compare the performance of the developed methodology to other existing methods.

At the end of year 1 and beginning of year 2, validate developed computational methods on domestic horse data collected thus far by comparing biological expert knowledge on leaders, communities, and critical individuals to those identified by computational tools. Use the zebra data to identify performance problems on noisy and lossy data from wild populations with limited human intervention.

Year 2: Continue collection of biological data on wild zebra populations in Kenya. Adjust the data collection protocols to meet the emerging requirements of the computational techniques. Follow closely the individuals identified as "critical" by computational methods and induce perturbations of the social structure. Start transfer of data collection responsibilities to Kenyan assistants.

Continue to develop and implement algorithms for dynamic social networks and focus on speed and scalability of those algorithms. Validate using the experimental setup on domestic horses and the initial data on zebra. Work on developing rigorous validation and accuracy measurement

techniques. Make a limited deployment of our software for other interested users to incorporate their feedbacks to improve the quality of our software. We will hold a *formal review* of our progress at the end of the year (see below).

Year 3: Data collection in year three will be done by field assistants in Kenya with assistance and oversight by PI Rubenstein. Targeted data will be collected to fine-tune and evaluate specific aspects of the methodology. By the end of the year, complete a full evaluation of the accuracy and robustness of our algorithms and software using the wild zebra populations at Mpala Research Centre, Kenya. Make our software and methodology freely available and will conduct limited user-testing of all computational tools. In terms of outreach, provide a user-friendly interface appropriate for users outside the immediate scientific community.

7.4 Dissemination

We will disseminate our research at various computer science and computational biology conferences and will publish both in biological, computer science, and social sciences journals to reach all scientific communities potentially interested in our research. We will make the software freely available on the web and will demonstrate our software at conferences. We will organize interdisciplinary workshops, bringing together computer scientists, mathematicians, biologists, and social and behavioral scientists.

7.5 Progress Review

We will hold a scientific review of progress and future directions at the end of second year of the project. We will host a one-day symposium centered on the research topics of our project. The symposium will be comprised of invited talks by outside experts as well as progress report talks by the members of this team. A panel discussion will be held with the outside experts and a written report will be generated. The composition of the project personnel will be also reviewed and new tasks and goals will be determined. We will hold another review of progress at the end of the final year. At the end of this review, we will produce a detailed report of successes of our projects.

References

[1] E. Airoldi and B. Malin. Data mining challenges for electronic safety: the case of fraudulent intent detection in e-mails. In *Proceedings of the Privacy and Security Aspects of Data Mining Workshop, in conjunction with the 4th IEEE Internation Conference on Data Mining*, Brighton, England, 2004.

[2] J. Aizen, D. Huttenlocher, J. Kleinberg, and A. Novak. Traffic-based feedback on the web. *Proceedings of the National Academy of Sciences*, 101(Suppl.1):5254–5260, 2004.

[3] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, 2002.

[4] J. Altmann. Observational study of behavior: sampling methods. *Behaviour*, 49:227–267, 1974.

[5] Z. Bar-Yossef, A. Broder, R. Kumar, and A. Tomkins. Sic transit gloria telae: Towards an understanding of the web's decay. In *Proceedings of WWW*, 2004.

[6] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, 2005.

[7] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614, August 2002. 10.1016/S0378-4371(02)00736-7.

[8] J. Baumes, M. Goldberg, M. Magdon-Ismail, and W. Wallace. Discovering hidden groups in communication networks. In *Proceedings of the 2nd NSF/NIJ Symposium on Intelligence and Security Informatics*, 2004.

[9] T. Berger-Wolf and J. Saia. A framework for analysis of dynamic social networks. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 523–528, 2006.

[10] T. Y. Berger-Wolf, B. Hart, and J. Saia. Discrete sensor placement problems in distribution networks. *Journal of Mathematical and Computer Modeling*, 2005.

[11] T. Y. Berger-Wolf, S. Sheikh, W. Chaovalitwongse, B. DasGupta, , and M. V. Ashley. Reconstructing sibling relationships from microsatellite data. Submitted.

[12] S. Bernstein, A. Clearwater, S. Hill, C. Perlich, , and F. Provost. Discovering knowledge from relational data extracted from business news. In *Proceedings of the Workshop on Multi-Relational Data Mining, in conjunction with the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

[13] I. Boman, C. Abdallah, E. Schamiloglu, and J. Saia. Self-healing algorithms for reconfigurable networks. In *To appear in International Symposium on Stabilization, Safety and Security of Distributed Systems (SSS)*, 2006.

[14] K. Börner, L. DallAsta, W. Ke, and A. Vespignani. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. In *Complexity, Special issue on Understanding Complex Systems*, 2006. in press.

[15] K. Börner, J. Maru, and R. Goldstone. The simultaneous evolution of author and paper networks. *PNAS*, 101(Suppl 1):5266–5273, 2004.

[16] R. Breiger, K. Carley, and P. Pattison, editors. *Dynamic Social Network Modeling and Analysis*. The National Academies Press, Washington, D.C., 2003.

[17] A. Broido and K. Claffy. Internet topology: connectivity of IP graphs. In *Proceedings of SPIE ITCom*, 2001.

[18] S. J. Cairns and S. J. Schwager. A comparison of association indices. *Animal Behaviour*, 35:1454–1469, 1987.

[19] K. Carley. Communicating new ideas: The potential impact of information and telecommunication technology. *Technology in Society*, 18(2):219–230, 1996.

[20] K. Carley. Dynamic network analysis. In R. Breiger, K. Carley, and P. Pattison, editors, *Dynamic Social Network Modeling and Analysis*, pages 133–145. The National Academic Press, Washington, D.C., 2003.

[21] K. Carley and M. Prietula, editors. *Computational Organization Theory*. Lawrence Erlbaum associates, Hillsdale, NJ, 2001.

[22] L. Chen and K. Carley. The impact of social networks in the propagation of computer viruses and countermeasures. *IEEE Trasactions on Systems, Man and Cybernetics*, forthcoming.

[23] D. Croft, J. Krause, and R. James. Social networks in the guppy (*poecilia reticulata*). *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 271:516–519, 2004.

[24] D. P. Croft, R. James, P. Thomas, C. Hathaway, D. Mawdsley, K. Laland, and J. Krause. Social structure and co-operative interactions in a wild population of guppies (*poecilia reticulata*). *Behavioural Ecology and Sociobiology*, In Press.

[25] D. P. Croft, R. James, A. J. W. Ward, M. S. Botham, D. Mawdsley, and J. Krause. Assortative interactions and social networks in fish. *Oecologia*, 143:211–219, 2005.

[26] P. C. Cross, J. O. Lloyd-Smith, and W. M. Getz. Disentangling association patterns in fission–fusion societies using african buffalo as an example. *Animal Behaviour*, 69:499–506, 2005.

[27] P. Desikan and J. Srivastava. Mining temporally evolving graphs. pages 13–22, New York, NY, USA, 2004. ACM Press.

[28] S. Eubank, H. Guclu, V. Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429:429:180–184., Nov 2004. Supplement material.

[29] S. Eubank, V. Kumar, M. Marathe, A. Srinivasan, and N. Wang. Structural and algorithmic aspects of massive social networks. *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 718–727, 2004.

[30] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *Proceedings of WWW*, 2003.

[31] A. Fiat, J. Saia, and M. Young. Making chord robust to byzantine attacks. In *Proceedings of the European Symposium on Algorithms(ESA)*, 2005.

[32] I. R. Fischhoff, S. R. Sundaresan, J. Cordingley, H. M. Larkin, M.-J. Sellier, and D. I. Rubenstein. Social relationships and reproductive state influence leadership roles in movements of plains zebra (*equus burchellii*). *Animal Behaviour*, 2006. Submitted.

[33] I. R. Fischhoff, S. R. Sundaresan, J. Cordingley, and D. I. Rubenstein. Habitat use and movements of plains zebra (*equus burchelli*) in response to predation danger from lions. Submitted.

[34] L. Freeman. Finding social groups: A meta-analysis of the southern women data. In R. Breiger, K. Carley, and P. Pattison, editors, *Dynamic Social Network Modeling and Analysis*. The National Academies Press, Washington, D.C., 2003.

[35] J. R. Ginsberg and T. P. Young. Measuring association between individuals or groups in behavioural studies. *Animal Behaviour,*, 44:377–379, 1992.

[36] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[37] M. Hack and D. I. Rubenstein. Zebra zones. *Natural History*, 107(2):26–29, 1998.

[38] M. A. Hack, R. East, and D. I. Rubenstein. Plains zebra (*equus burchelli* gray). In P. D. Moehlman, editor, *Equids: Zebras, Asses, and Horses. Status Survey and Conservation Action Plan*, pages 43–57. IUCN/SSC Equid Specialist Group, IUCN, Switzerland and Cambridge, UK, 2002.

[39] J. Hipp, U. Güntzer, and G. Nakhaeizadeh. Algorithms for association rule mining — a general survey and comparison. *SIGKDD Explorations*, 2(1):58–64, July 2000.

[40] N. Hummon and K. Carley. Social networks: As normal science. *Social Networks*, 15:71–106, 1993.

[41] P. Jaccard. The distribution of flora in the alpine zone. *The New Phytologist*, 11(2):37–50, 1912.

[42] P. Juang, H. Oki, Y. Wang, M. Martonosi, L. Peh, and D. Rubenstein. Energy-efficient computing for wildlife tracking: Design tradeoffs and early experiences with zebranet. In *ASPLOS, San Jose, CA*, Oct. 2002.

[43] M. Keeling. The effects of local spatial structure on epidemiological invasions. *Proc. R. Soc. Lond. B*, 266:859–867, 1999.

[44] D. Kempe, J. Kleinberg, and A. Kumar. Connectivity and inference problems for temporal networks. *J. Comput. Syst. Sci.*, 64(4):820–842, 2002. doi 10.1006/jcss.2002.1829.

[45] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[46] V. Kin, S. Lewis, J. Saia, and M. Young. Choosing a Random Peer in Chord. *To appear in Algorithmica*, 2006.

[47] V. King and J. Saia. Choosing a random peer. In *Proceedings of the Twenty-Third Annual ACM Symposium on Principles of Distributed Computing (PODC)*, 2004.

[48] V. King, J. Saia, V. Sanwalani, and E. Vee. Scalable leader election. In *Proceedings of the Symposium on Discrete Algorithms(SODA)*, 2006.

[49] V. King, J. Saia, V. Sanwalani, and E. Vee. Towards secure and scalable computation in peer-to-peer networks. In *To Appear in Foundations of Computer Science(FOCS)*, 2006.

[50] J. Kleinberg. Temporal dynamics of on-line information streams. Draft chapter for the forthcoming book Data Stream Management: Processing High-Speed Data Streams (M. Garofalakis, J. Gehrke, R. Rastogi, eds.), Springer.

[51] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the Thirty-second ACM Symposium on Theory of Computing (STOC)*, 2000.

[52] J. Kleinberg. Small-world phenomena and the dynamics of information. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence,*, Morgan Kaufman, 2001.

[53] J. Kleinberg. The small-world phenomenon and decentralized search. *SIAM News*, 37, 2004.

[54] W. Koehler. A longitudinal study of web pages continued: a consideration of document persistence. *Information Research*, 9(2):paper 174, 2004. [Available at http://InformationR.net/ir/9-2/paper174.html].

[55] G. Kolata. Ideas and trends; enron offers an unlikely boost to e-mail surveillance. *New York Times*, May 22 2005.

[56] M. Kretzschmar and M. Morris. Measures of concurrency in networks and the spread of infectious disease. *Math. Biosci.*, 133:165–195, 1996.

[57] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proc. International WWW Conference*, 2003.

[58] M. Lahiri and T. Y. Berger-Wolf. Structure prediction in temporal networks using frequent subgraphs. Submitted.

[59] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2005.

[60] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. *Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, 2003.

[61] D. Lusseau and M. E. J. Newman. Identifying the role that individual animals play in their social network. *Proc. R. Soc. London B (Suppl.)*, 271:S477–S481, 2004.

[62] M. Magdon-Ismail, M. Goldberg, W. Wallace, and D. Siebecker. Locating hidden groups in communication networks using hidden markov models. In *Proceedings of the International Conference on Intelligence and Security Informatics (ISI 2003)*, Tucson, AZ., 2003.

[63] B. Malin. Data and collocation surveillance through location access patterns. In *Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference*, Pittsburgh, PA, June 2004.

[64] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovery of frequent episodes in event sequences. *Data Min. Knowl. Discov.*, 1(2):259–289, 1997.

[65] L. A. Meyers, M. Newman, and B. Pourbohloul. Predicting epidemics on directed contact networks. *Journal of Theoretical Biology*, 240:400–418, 2006.

[66] L. A. Meyers, B. Pourbohloul, M. Newman, D. Skowronski, and R. Brunham. Network theory and sars: Predicting outbreak diversity. *Journal of Theoretical Biology*, 232:71–81, 2005.

[67] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):25102, 2001.

[68] M. E. J. Newman. From the Cover: The structure of scientific collaboration networks. *PNAS*, 98(2):404–409, 2001.

[69] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[70] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev.*, 69, 2004.

[71] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev.*, 69, 2004.

[72] A. Ntoulas, J. Cho, and C. Olston. What's new on the web? The evolution of the web from a search engine perspective. In *Proceedings of WWW*, 2004.

[73] C. Papadimitriou. Computational aspects of organization theory. *Lecture Notes in Computer Science*, 1997.

[74] C. Papadimitriou and E. Servan-Schreiber. he origins of the deadline: Optimizing communication in organizations. *Complexity in Economics.*, 1999.

[75] R. Pastor-Satorras and A. Vespignani. *Evolution and structure of the Internet.* Cambridge University Press, Cambridge, 2004.

[76] J. J. Ramasco, S. N. Dorogovtsev, and R. Pastor-Satorras. Self-organization of collaboration networks. *Physical Review E*, 70:036106, 2004.

[77] J. M. Read and M. J. Keeling. Disease evolution on networks: the role of contact structure. *Proc. R. Soc. Lond. B*, 270:699–708, 2003. doi 10.1098/rspb.2002.2305.

[78] Y. Ropert-Coudert and R. Wilson. Trends and perspectives in animal-attached remote sensing. *Fronteirs in Ecology and Environment*, 3:437–444, 2005.

[79] D. I. Rubenstein. Herd dynamics: Why aggregations vary in size and complexity. In M. Bekoff, editor, *Encyclopedia of Animal Behavior*, pages 994–1000. Greenwood Press, Westport, CT, 2004.

[80] D. I. Rubenstein. Zebra sociality: Different stripes for different types. *Balliol College Record*, pages 18–24, 2004.

[81] D. I. Rubenstein and M. Hack. Natural and sexual selection and the evolution of multi-level societies: insights from zebras with comparisons to primates. In P. Kappeler and C. P. van Schaik, editors, *Sexual Selection in Primates: New and Comparative Perspectives*, pages 266–279. Cambridge University Press, 2004.

[82] D. I. Rubenstein, S. Sundaresan, I. Fischhoff, and D. Saltz. Social networks in wild asses: Comparing patterns and processes among populations. In A. Stubbe, P. Kaczensky, R. Samjaa, K. Wesche, and M. Stubbe, editors, *Exploration into the Biological Resources of Mongolia*, volume 10. Martin-Luther-University Halle-Wittenberg, 2007. In press.

[83] D. R. Rubenstein, D. I. Rubenstein, P. W. Sherman, and T. A. Gavin. Pleistocene Park: Does re-wilding North America represent sound conservation for the 21st century? *Biological Conservation*, 132:232–238, 2006.

[84] J. B. Silk, S. Alberts, and J. Altmann. Social bonds of female baboons enhance infant survival. *Science*, 302:1231–1234, 2003.

[85] T. Snijders. The Statistical Evaluation of Social Network Dynamics. *Sociological Methodology*, 31(1):361–395, 2001.

[86] D. Spinellis. The decay and failures of web references. *Communications of the ACM*, 46:71–77, 2003.

[87] S. R. Sundaresan, I. R. Fischhoff, J. Dushoff, and D. I. Rubenstein. Network metrics reveal differences in social organization between two fission-fusion species, Grevy's zebra and onager. *Oecologia*, 2006. doi 10.1007/s00442-006-0553-6.

[88] W. Tong, D. I. Rubenstein, and B. Shapiro. On the genetic structure of plains zebra (*equus burchelli*) populations: insights from noninvasive microsatellite genotyping. Submitted.

[89] M. Tsvetovat, K. Sycara, Y. Chen, and J. Ying. Customer coalitions in electronic marketplaces. In U. C. Frank Dignum, editor, *Agent-Mediated Electronic Commerce III, Lecture Notes on Artificial Intelligence*. Springer-Verlag, 2003.

[90] J. Tyler, D. Wilkinson, and B. Huberman. Email as spectroscopy: Automated discovery of community structure within organizations. In *Proceedings of the First International Conference on Communities and Technologies*, 2003.

[91] S. Wasserman and F. K. *Social Network Analysis.* Cambridge University Press, Cambridge, MA, 1994.

[92] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[93] J. Weinstock, E. Willerslev, A. Ster, W. Tong, S. Ho, D. I. Rubenstein, J. Stoprer, J. Barnes, L. Martin, C. Brovi, A. Preito, D. Froese, E. Scott, L. Xulong, and A. Cooper. Evolution, systematics, and phylogeography of pleistocene horses in the new world: A molecular perspective. *PloS Biology*, 3(8):1373–1379. e241, August 2005.

[94] B. Wellman. An electronic group is virtually a social network. In S. Kiesler, editor, *Culture of the Internet*, pages 179–205. Lawrence Erlbaum, Mahwah, NJ, 1997.

[95] H. Whitehead and S. DuFault. Techniques for analyzing vertebrate social structure using identified individuals: review and recommendations. *Advances in the Study of Behavior*, 28:33–74, 1999.

[96] W. Wickler. *Mimicry in Plants and Animals*. McGraw-Hill, New York, NY, 1968.

[97] Q. Yang and X. Wu. 10 challenging problems in data mining research. IEEE International Conference on Data Mining (ICDM) Presentation Slides, 2005. KDnuggets : News : 2005 : n24 : item3.

[98] P. Zhang, C. Sadler, T. Liu, I. Fischhoff, M. Martonosi, S. A. Lyons, and D. I. Rubenstein. Habitat monitoring with ZebraNet: Design and experiences. In N. Bulusu and S. Jha, editors, *Wireless Sensor Networks: A Systems Perspective*, pages 235–257. Artech House, Norwood, MA, 2005.