From Market Baskets to Mole Rats:
Using Data Mining Techniques to Analyze RFID
Data Describing Mole Rat Behavior

Dan P. McCloskey1,2, Michael E. Kress3,7, Susan P. Imberman4,7, Igor Kushnir5, Susan Briffa-Mirabella2,6
1. Dept. Psychology 2. CSI/IBR Center for Developmental Neuroscience, 3. Vice President of Technology Systems,
4. Dept. Computer Science 5. Masters Program in Neuroscience, 6. Program in Neuroscience, Department of Biology,
The Graduate Center at CUNY 7. Doctoral Program Computer Science the Graduate Center at CUNY
College of Staten Island, 2800 Victory Blvd, Staten Island, NY 10314
Correspondence to: susan.imberman@csi.cuny.edu

ABSTRACT
The use of new technologies, such as RFID sensors, provides scientists with novel ways of doing experimental research. As scientists become more technologically savvy and use these techniques, the traditional approaches to data analysis fail given the huge amounts of data produced by these methods. In this paper we describe an experiment in which colonies of naked mole rats were tagged with RFID transponders. RFID sensors were strategically placed in the mole rat caging system. The goal of this experiment was to document and analyze the interactions between animals. The huge amount of data produced by the sensors was not analyzable using the traditional methods employed by behavioral neuroscience researchers. Computational methods used by data miners, such as cluster analysis, association rule mining, and graphical models, were able to scale to the data and produce knowledge and insight that was previously unknown. This paper describes in detail the experimental setup and the computational methods used.

Categories and Subject Descriptors
J. Computer Applications J.4 SOCIAL AND BEHAVIORAL SCIENCES Psychology
H. Information Systems H.2 DATABASE MANAGEMENT
H.2.8 Database applications Data Mining

General Terms
Algorithms, Measurement, Experimentation,

Keywords
cluster analysis, association rule mining, maximal frequent itemsets, visualization, RFID

1. INTRODUCTION
The study of animal behavior can take place in a variety of settings. Animals can be observed in their natural environment (Field Studies) and data collected can be as low-tech as birdwatchers recording data from banded birds who land in their back yard, or as high tech as real-time GPS tracking of individual animals as they migrate thousands of miles. In the laboratory setting, the study of animal behavior has been considerably low tech.

Most laboratory measurements of animal behavior involve animals (typically rats or mice) that are placed in a maze environment while an experimenter quantifies behavioral events (e.g. maze choices) either directly, or through video recording. This approach has been questioned for its external validity, as brain and behavior have been shown to be quite sensitive to environment and experience. Therefore, removing animals from their home environment may cause inaccurate measurement of behavior (See [27] for discussion).

An alternative approach to removing animals from their housing environments and measuring behavior in a novel and artificial setting, is to capture behavior in the housing environment. This approach has the advantage of limiting animal stress and behavioral response to the artificiality of the testing experience. It also allows for the analysis of behavior across multiple time scales (seconds to weeks). Recent developments in automated video-oriented animal behavior analysis have allowed for the advanced detection of abnormalities in transgenic mouse models [14]. However these behavior detection systems have the limitation that animals are housed individually, which can be stressful for most rodents [26].

An area of growing interest in behavioral neuroscience is the study of social behaviors, with respect to how the brain is modified in various conditions such as autism [10]. While social approach behavior (investigating a caged unfamiliar animal) can be automated in an artificial maze environment [11], no studies to date have attempted to measure social behavior over long periods of time in the animal’s own housing environment. Studying animals in this way may be important if the processes that drive social behavior in a novel location with an unfamiliar animal, differ from those that drive social behavior in a familiar setting with familiar animals. Traditionally, behavioral scientists use automated video systems, but these systems are incapable of distinguishing multiple laboratory animals, particularly over long periods of time in a complex housing environment.

Our research in this area therefore required a system that was able to deliver a snapshot of the location of multiple animals in a complex housing environment. While this system should be useful for the study of both rats and mice, we wanted to develop a system that could also measure activity in a specific laboratory animal, the naked mole rat. Naked mole rats (NMR) are highly...
social rodents with colonies that can grow to 75 in a single housing environment. We were interested in knowing the locations of each animal, with a sub second time resolution, over a period of days.

Our solution to this problem was to quantify social behavior using the premise that animals that spend more time in the presence of other animals are more social than animals that spend more time alone. To quantitatively analyze these behaviors, an RFID (radio frequency identification) system, consisting of transponders and readers, was used. RFID provides automatic identification of physical objects using radio waves. Because each RFID transponder has a unique ID, an unlimited number of animals could be tracked within the same housing environment. Our initial question was, “Which mole rats socially interacted with each other?” To answer this, we developed a system based on the methods reported by Kritzler and colleagues to track mouse behavior [17,18,19]. The underlying premise was that mole rats tagged with transponders that socialized with each other, would be located in the proximity of the same reader. The major problem encountered was that the data generated by this system was too massive to be analyzed using the traditional methods by behavioral scientists.

The organization of this paper outlines our approach as follows: Section 2 outlines the experimental design and the software used for data capture. Section 3 details the data mining algorithms used for data analysis and the results. Section 4 discusses our conclusions and we summarize in section 5.

2. EXPERIMENTAL DESIGN

2.1 Animal Model

The naked mole rat (Heterocephalus Glaber) is one of 18 species of African mole-rats, which belong to the hystricognath suborder of rodents. All African mole-rats, the naked mole rats inhabit some of the harshest environmental habitats in the arid regions of central and eastern Ethiopia, central Somalia, and Kenya. Naked mole-rats are highly successful in the wild, and are in no danger of extinction at this time. They have adapted by evolving skills to work together, including skills for social tolerance and cooperation in large groups. They have the largest colony size of any social vertebrate, which averages about 75 animals in the wild, and the largest litter size of any mammal [4,25]. The colony is maintained through a cooperative breeding system, with fewer than 5% of the animals achieving reproductive status in their lifetime. These animals are one of only two mammalian species (both African mole rats) with a eusocial reproductive hierarchy similar to that of bees, termites, or ants. The non-breeding animals cooperate in caring for young (retrieval, licking, grooming, and provision of food), and protecting and maintaining the colony [20]. Cooperative behaviors and social activity among subordinate animals allows for the study of social behaviors that do not appear to be confounded by mating behavior or the goal of procreation, and permits shared housing of dozens of animals in the same environment. Naked mole rats are exclusively subterranean and maintain an extensive burrow system, with 3-4 km of tunnels, used to identify tuber food sources. The animals also maintain communal nest and toilet chambers that are areas of high social interaction. Figure 1 shows an example of two naked mole rats in their laboratory housing environment.

While previous studies have examined interactions of pairs of animals temporarily removed from their housing environment [9], or activity patterns of animals living in isolation [23], the present study was designed to measure activity and social behavior in the intact, unperturbed colony. Therefore, the results from these analyses would provide the first evidence of group behavior dynamics of a naked mole rat colony. It was hypothesized that within each colony of naked mole rats, there would be measurable differences in the level of social behavior.

2.2 RFID Tracking in home environment

Each naked mole rat in the animal laboratory was implanted subcutaneously with a Trovan Unique radio frequency identification transponder (transponder size 11.5 x 2.2 mm; MicrochipID Lake Zurich, IL) in March 2010. The transponders are passive and therefore do not require batteries, and should not need to be replaced during the lifespan of the animal (over 28 years in captivity, the longest living rodent on Earth; [5]). In each colony, animals are housed in 3 standard mouse tub cages connected by over 7 meters of clear polycarbonate tubing (50.8 mm inner diameter) with stationary circular RFID antennae (100 mm inner diameter) placed around the tubing at multiple locations, and connected to a backend computer (See Figure 2).

The dataset used for the current analyses were collected from 3 separate colonies of naked mole rats (n = 16, 22, and 38 animals respectively). Colonies were physically separated from each other, so the analyses should show no interaction between animals of different colonies, thus validating the experimental model and the analytical techniques.

The RFID tracking system consists of a tag, a reader, and a backend computer (See Figure 3). Below is a list of each component along with a description of that component.

![Figure 1: Image of 2 adult naked mole rats from the Degas colony traveling into the nest chamber after passing through antenna number 4. The animal to the rear is the queen of the colony, the only reproductively active female.](image1)

![Figure 2: Arrangement of Degas colony semi-natural environment with chambers (toilet chamber on bottom right) connecting tunnels, antennae, and the backend computer for compiling data.](image2)
The event in the text file included the animal ID (unique 10 digit alphanumeric code), time of entry, and reader number (1-10). Text files containing 24 hours of data entries were parsed using Matlab, by location (antenna) and a state matrix number (1-10). Text files containing 24 hours of data entries were parsed using Matlab, by location (antenna) and a state matrix identifying the last known location for each animal was processed for each event.

3. DATA METHODOLOGY AND RESULTS

As mentioned above, the goal of the current study was to measure animal social behavior. Traditional approaches for this type of measurement involve recording the amount of time a test animal spends in the area of a maze where another animal resides during a 15 minute test period [11]. With a total number of dyads in a group equal to N(N-1)/2, traditional social approach behavior analysis for a colony of 33 animals would require 7,920 minutes of behavioral observation.

The alternative approach is to monitor the amount of time animals spend in each possible dyad at specific locations in the housing environment. However, this process would require mining or summarizing very large datasets. We present here several successful methods in data mining and summary data analysis.

3.1 Cluster Analysis

Hierarchical and K-means clustering techniques were implemented with the goals of organizing or partitioning the data, respectively, such that animals that spend time in the same location are homogenous and animals with different location preferences are well-separated. Clusters were determined using iterative functions in Matlab based on the state matrix dataset. Figure 4 shows a histogram organized by the results of a K-means cluster analysis, with the bars representing the normalized frequency of “visits” (y axis) by each animal (1-76, x axis) to each reader location (1-10, z axis).

This type of analysis is beneficial as it utilizes the entire dataset, including every movement of each animal during the 24-hour recording period. It also is a useful tool to identify animals with similar “place preferences”, demonstrated here by the separation of the three colonies into different clusters. However, although two animals may have the same place preference, these clustering techniques will not indicate whether two animals were in the same place at that same time. Therefore, while we found clustering useful for identifying common and unusual patterns of behavior, we were limited in the conclusions we could draw regarding social behavior.

3.2 Adjacency Matrix Sampling

The second approach was to generate a consensus structure of the network organization using multiple adjacency matrices. Adjacency matrices are n x n matrices constructed such that for time T, if animals i and j are indicated at the same reader location then the i,jth cell is equal to 1 and if animal i is in a different location then both the i,jth and j,i th cells are equal to 0. The adjacency matrix is undirected and symmetrical such that for adjacency matrix A: A_{ij} = A_{ji}.

Adjacency matrices can be easily graphed using statistical (R) or graphical (Gephi) software, where the vertices (nodes) are represented by polygons, and the edges are represented by lines connecting a node. This type of graphical representation of social relationship, often referred to as a sociogram, is common in the field of social network analysis [9,25]. Thus, in the example above, the sociogram would show nodes representing animals i,j and k, with lines connecting nodes i and j, but with node k not connected to either i or j.

Clearly, choosing a single time-point may not be a good measure of social behavior, as animals i and j may be merely passing each other at the same location at time T, and not interacting at all. Therefore, a more reliable indicator of social behavior would require multiple time points sampled during the 24 hour period. To this end, we created adjacency matrices at 100 random time points during the 24 hour data acquisition period, and created a consensus matrix using the centralgraph routine in the sna package in R [7]. This function calculates the mean dichotomized adjacency matrix of the 100 sample matrices.
Figure 5: Sociogram computed in R using the centralgraph function representing the average interactions from 100 randomly sampled adjacency matrices from the data used in Figure 4. Figure 6: Gephi file sociogram based on maximal frequent itemsets from the same datafile shown in Figure 5. Notice that the three colonies are once again separated, and that the colony identity of some animals in Figure 5 has been resolved.

One difficulty in this type of analysis, however, is deciding what the appropriate sample size is to represent the population of social interactions. Here, we used 100 random sampled time points over a 24-hour period, but without any scientific basis for the number of samples or length of the recording period. To avoid this problem, a scalable data mining procedure, which is not prone to sampling bias, may be advantageous.

3.3 Association Rule Mining

Association rule mining has been used in market basket analysis for finding groups of items that are purchased together. Each store transaction corresponds to groups of items in a customer's market basket. This results in sets of items that are frequently purchased together, along with rules to the effect that, "customers that buy cereal are likely to buy milk". These rules hold with minimum statistical significance and confidence. Our goal for this study was to find mole rats that group together, given readings captured from a reader. There seemed a natural mapping from the traditional association rule market basket problem to the groupings defined by the readings from the RFID system. One can think of the readers as corresponding to the cash register scanner in a store, the groups of mole rats captured at each reader corresponding to a market basket transaction, and the results yielding frequent groups of mole rats that socialize together, along with rules to the effect of, "when animals 5 and 6 are together we are likely to see animal 8 as well".

More formally, given a set $I = \{ i_1, i_2, i_3, ..., i_d \}$ of items, a transaction is a subset $X$ of $I$. A Subset of the items in a transaction is also a subset of $I$ and is called an itemset. A frequent itemset is a subset of $I$ that satisfies some minimum percentage of transactions. This value is known as the support of the itemset. Support is an indication of the itemset's statistical significance. Let $X$ and $Y$ be subsets of $I$ such that $X \cap Y = \phi$. An association rule is a probabilistic implication $X \Rightarrow Y$. Given an association rule $X \Rightarrow Y$ as defined above, the support of the rule is the support of the itemset defined by $X \cup Y$. The confidence of a rule is the number of transactions that satisfy the itemset defined by $X \cup Y$ divided by the number of transactions that satisfy $X$. A rule that has support above a user defined minimum called minsup and confidence above a user defined minimum called minconf is an interesting association rule [1,2,3].

Recently there has been a focus on mining sets of itemsets that are representative of the set of frequent itemsets. Mining frequent itemsets can become cost prohibitive in dense datasets with long patterns of 20 items or more. For these types of datasets research has focused on mining closed frequent itemsets [21,22,24] and maximal itemsets [6,13]. Given a frequent itemset $X$, $X$ is a closed frequent itemset if there exists no $Y \supseteq X$ such that the support of $Y$ equals the support of $X$. A frequent set $X$ is a maximal frequent itemset if for all supersets $Y$ of set $X$, set $Y$ is infrequent. [13] The frequent itemsets are a superset of the closed frequent itemsets, which in turn are a superset of the maximal frequent itemsets.

Since we were looking to find sets of animals that socialize, it made sense to look at the maximal frequent itemsets rather than the rules these itemsets defined. For instance, rat5 $\Rightarrow$ rat24 has no different confidence measure than rat24 $\Rightarrow$ rat5 since both rat24, and rat5 would trigger the same reader at the same time if they collocated. We chose to find the maximal frequent itemsets since these coded for a less redundant set of items. [28]

With direction from the domain expert, we found all maximal itemsets with a minimum support of 40%, the interpretation being, find all groups of mice that socialize at least 40% of the time. For the dataset containing 76 mole rats, 369 maximal itemsets were...
found. Klemettinen et al. [16] describe a method of visualizing association rules using a graphical representation. Attributes are represented by nodes, and the directed arcs between these represent the rules. Building on this, we represented the maximal itemsets generated for the naked mole rat domain by an undirected graph. Given that each animal is represented by a node, the itemsets code for undirected edges between animals within the same itemset. Itemsets can be graphically connected to each other via common items. For example, given maximal itemset \{6, 25, 29\} and maximal itemset \{6, 31\} we can see that animal 6 associates with animals 25, 29, and 31. Figure 6 shows the undirected graph found from the maximal frequent itemsets. Note that the results are generally similar to those in Figure 5, but that association rule mining resolves some issues, such as the colony identity of animals 12, 40, and 68. Animal 57 lived in the same colony as animals 7, 18, 4 (etc) but the co-localization of this animal with colony mates was too infrequent to create an edge in either the adjacency matrix or association rule mining analyses.

4. CONCLUSIONS
The measurement of the social behavior of laboratory animals has become more in demand in recent years as investigators seek to understand the neural mechanisms of autism and schizophrenia [10]. Traditional methods used to study social behavior lack external validity as the testing environment is artificial and may be confounded by the stress of encounters with unfamiliar animals. These methods are also time consuming and not well-suited for the study of large groups of animals.

The current study used multiple complementary approaches to measure the social organization of large groups of RFID tracked laboratory animals in the home environment. This approach is advantageous, as it improves external validity, and can accommodate large groups. However, the very large datasets created by RFID tracking pose new problems for researchers in the field of behavioral neuroscience.

Multiple solutions were attempted to solve the problem of drawing task-relevant conclusions from these large data sets. First, hierarchical and K-series mean clustering techniques were used to demonstrate groups of animals with similar location preferences. While this approach is useful to demonstrate patterns of behavior, it is not conclusive whether animals shared the same space at the same time.

Our next approach was to create a central graph based on multiple (>100) animal-location adjacency matrices created from random time points throughout the day. This technique is useful for identifying which animals share the same space, and is amenable to social network graphing and analysis (i.e. degree of centrality), but is problematic in that large preliminary datasets must be analyzed before assessing how much sampling is enough to accurately depict the true social relationships.

Our final approach, the association rule mining approach, is advantageous because it uses the whole dataset and thus avoids the issues of sampling bias or inadequacy. As shown here, the output of this analysis is also amenable to graphing in social network analysis programs and analysis of social network statistics.

Together, these results show, for the first time that despite having nearly identical genetic backgrounds, siblings in a naked mole rat colony show measurable differences in the level of social behavior. This validates the use of this technology as a new approach to measuring social behavior in the laboratory animal.

5. SUMMARY
Jim Gray of Microsoft Research defined eScience as the intersection between IT and scientists. He goes on to define a fourth paradigm of scientific inquiry as being "data intensive science". The work proposed in this paper is a prime example of the kinds of results one can obtain when pairing scientists with computer scientists.

This work provides a template for using automated data collection tools to measure social behavior in large groups. The experiments demonstrated here were "proof-of-concept" in nature to confirm that animals housed in physically separated environments would appear as separate social networks in our analyses, but they also yielded information about the differing levels of social behavior within each colony.

For future work, we are recording single colonies with higher spatial resolution (~25 cm inter-reader distances) and have achieved even greater understanding of the role of individual animals in the larger social network. These methods are also being applied to a more traditional laboratory animal species, the Long Evans rat, with similar success.

Because of the temporal nature of the data, we will investigate the application of additional data mining techniques such as sequential temporal data analysis, and social network analysis to this experimental dataset. These analytical methods allow us to answer questions such as, "How long are animal groups associating with each other?, Do social groups congregate only in one area or do they socialize in other maze areas as well?, and Which animal is initiating a social interaction?" The technology and analytical tools provides a breadth of possibilities in this area, opening up the investigation of behaviors that were too difficult to be previously analyzed.

These analyses yielded informative results because they allowed for the measurement of multiple animals simultaneously. To our knowledge, this social network approach has not been used previously to identify group behavior dynamics in laboratory animals. Based on this combination of methods, we intend to compare brain features of more social and less social animals from the same colony, with a long-term goal to understand neural correlates of social behavior deficits.

6. ACKNOWLEDGMENTS
This research was supported, in part, under National Science Foundation Grants CNS-0958379 and CNS-0855217, PSC-CUNY grant 62323-00-40, the Dormitory Authority of the State of New York, and the City University of New York High Performance Computing Center. We thank Dr. Bruce Goldman for generous donation of the naked mole rat colonies and Dr. Xiowen Zhang for advice on designing the RFID tracking system.

7. REFERENCES
introduction.


