Using Frequent Pattern Mining to Identify Behaviors in a Naked Mole Rat Colony

Susan P. Imberman, Ph.D.[4,5], Michael E. Kress, Ph.D.[3,5]. Dan P. McCloskey Ph.D.[1,2]

1.Dept. Psychology 2. CSI/IBR Center for Developmental Neuroscience, 3. Vice President of Technology Systems, 4. Dept. Computer Science
College of Staten Island, 2800 Victory Blvd, Staten Island, NY 10314, 5. Doctoral Program Computer Science the Graduate Center at CUNY,
Correspondence to: susan.imberman@csi.cuny.edu

Abstract

Animal behavior analysis has, in the past, taken a very low tech approach, with direct observer surveillance and automated video surveillance as the norm. These methods are insufficient when one wants to study interactions between large numbers of animals in their housing environment. In this paper we use a housing environment that has been equipped with a system of RFID sensors. RFID transponders were implanted into the study animal, the naked mole rat. The resulting data was analyzed using principal component analysis and frequent pattern mining. Results showed that these methods can identify time periods of high behavioral activity from that of low activity, along with which groups of animals interacted with one another.

Introduction

The study of animal social interactions is of interest to social psychologists so that we can better understand the various social disorders that affect humans Studies range from understanding the role of social behavior in mate selection and species fitness, to measuring brain correlates of social behavior.

Often, observing social interactions among animals requires removing animals from the larger group and observing them in pairs (Clarke and Faulkes 1999). This becomes tedious and time consuming when large groups of animals need to be observed and results may be influenced by the artificial nature of this type of observation.

Alternately one can observe animals in their home environment. This can be done using direct observation or by automated video surveillance systems. Observing large colonies of animals, such as those seen in the naked mole rat (NMR), can require the observation of 75 plus animals at any given time. Automated video systems fail because they are incapable of distinguishing among multiple laboratory animals over long periods of time in a complex housing environment. The problem then becomes, how can one observe social interactions between many animals simultaneously without removing them from their housing environment? Our solution was to use an RFID (radio frequency identification) system, consisting of transponders and readers. RFID provides automatic identification of physical objects using radio waves.

Transponders were embedded into each subject animal. Readers were judiciously placed in the housing environment. Each RFID transponder has a unique ID, enabling the ability to track individual animals through the readers. The organization of our system was based on the methods reported by Kritzler and colleagues to track mouse behavior (Kritzler, Lewejohann, and Krüger 2007) (Kritzler et. al. 2008) (Lewejohann et. al. 2009).

In previous work we demonstrated how social interactions can be measured by tracking dozens of NMRs simultaneously (McCloskey et al. 2011). A major problem encountered with this method was that our RFID system generated massive amounts of data which was not easily analyzed using the traditional methods employed by behavioral scientists. To remedy this, computational techniques such as those used in data mining, specifically principal components analysis (PCA) and frequent pattern mining, were used to identify behavioral patterns. This paper extends the previous work by attempting to characterize the social behavior patterns of NMRs with specific attention to how social behaviors are determined by the activity patterns of the larger colony. We asked whether social bonds between animals remain constant for the entire day, or whether different time points and behavior patterns cause animals to change the animals they socialize with. Understanding these types of social behaviors can inform us about areas and processes that may lead to better understanding, treatment, and prevention of social disorders such as schizophrenia and autism.

This paper is organized as follows. First we give background information on our study animal, the naked mole rat, and our computational methods. We then discuss our data set and how we applied PCA and frequent pattern analysis to this data. We end with our conclusions, summary and future directions.

Background

Naked mole rats represent a unique model system for the study of group social dynamics because they are the only mammalian species that use cooperative breeding (like honeybees or termites). Therefore they allow for the cohabitation of large groups of animals which can grow to upwards of 300 animals (Brett 1991) and live as long as 30 years (Buffenstein, 2008). Because all animals in a colony, except for the breeding pair, are siblings, we can study how the brain changes wrt social behavior without the confounds of stark genetic differences, or aggressive or breeding behaviors. Non-breeding animals cooperate in caring for young (retrieval, licking, grooming, and provision of food), and protecting and maintaining the colony (O'Riain 2000). NMRs are exclusively subterranean and maintain an extensive burrow system, with upwards of 3-4 km of tunnels, which are used to find tuber food sources. Within the lab housing environment animals maintain communal nest and toilet chambers which are areas of high social interaction.

Principal Component Analysis (PCA) is commonly used to find patterns in highly dimensional data. It does this by projecting the original data onto a lower dimensional space as defined by the eigenvalues and eigenvectors of the covariance matrix (Martinez, Martinez 2002). Each animal's social behavior defines a single dimension in the data space. Given the large numbers of animals, applying PCA to this highly dimensional space allowed us to project these behaviors onto a lower dimensional space, making it easier to see high and low time periods of social interaction. Thus we can find patterns defined by a small number of principle components which represent most of the observed data.

Frequent pattern mining has been used in the data mining community to find relationships between sets of items. Most notably this technique has been used by retailers to find groups of items purchased together by retail customers. Various implementations of frequent pattern mining have been developed over the years, the most notable of which is the Apriori algorithm (Agrawal and Srikant 1994) (Agrawal et. al. 1996). Although primarily used for market basket data, frequent pattern mining, at its simplest, can show us which variables or items co-exist with a minimum frequency. When applied to the RFID data generated by our system, frequent pattern mining was able to give insights into which NMR socialized together.

Data Methods and Results

In our laboratory setup, 33 NMRs were housed in standard mouse tub cages connected by over 7 meters of clear polycarbonate tubing (50.8 mm inner diameter). Each NMR was implanted subcutaneously with a Trovan Unique radio frequency identification transponder (transponder size 11.5 x 2.2mm; MicrochipID Lake Zurich, IL) in March 2010. Stationary circular RFID antennae (Trovan LID 650 readers (MicrochipID, Lake Zurich, 100 mm inner diameter) was placed around the polycarbonate tubing at multiple locations, and connected to a backend computer for data collection. Each time a tagged NMR passed through a sensor, a text file was updated with the animal ID (unique 10 digit alphanumeric code), time of entry, and reader number (1-14). Text files containing 24 hours of data entries were parsed using Matlab and a state matrix identifying the last known location for each animal was processed for each event. The process is event driven, with data generated only when the NMR moves through a sensor region. Figure 1 is a picture of the housing environment.

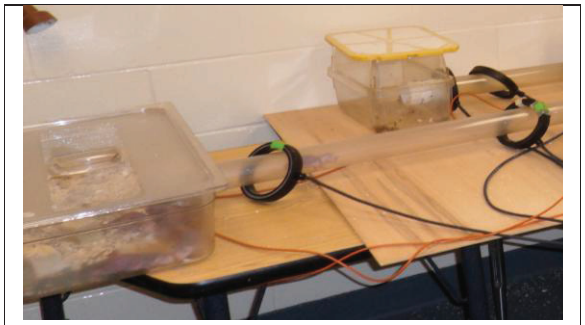

Figure 1: NMR housing environment showing mouse tubs and placement of RFID sensors.

The amount of data produced was large, with approximately 15,000 events recorded each hour, resulting in 3.1 million observations recorded over a 9 day period. The raw data was transformed into a state matrix as described above. Each row in the state matrix corresponded to an observation in the raw dataset. Columns were labeled with each NMR's ID. Each cell contained the data value of the current sensor location of the NMR at the time of the observation. In creating this matrix, we made the assumption, that NMRs that have not triggered a sensor reading, remained at their previous location. The state matrix formed the basis for the analyses using PCA and Apriori. Data was taken over a 24 hour period and was partitioned into 24 separate one hour time windows.

We calculated the covariance of the state matrix and analyzed this using PCA based on a singular value decomposition, resulting in the variance for each time window. The percent cumulative variance was then used to identify windows with different characteristic patterns. This analysis provided an understanding of hourly behavior patterns. Since each individual animal represented a single dimension in the data space, and the number of animals, hence dimensions started at 33, the hope was that PCA would be able to compress this highly dimensional set of behaviors into fewer behaviors in a lower dimensional space. In this case PCA could find data patterns of animal behavior (values/reader locations of variables/animals) by projecting the data into a new dimensionally lower space. The thought is that the cumulative relative variance will approach 1 quickly, using just a few of the largest principle components, i.e. the principle components which represent the patterns which show up most frequently in the data.

We found that most of the behavior can be projected onto three principal components. Given this, PCA results were able to identify two windows where behavior was markedly different: window 3 (2-3 AM) and window 14 (1-2 PM). We used these two windows to focus frequent pattern analysis on. Our goal was to show that NMR interactions were dependent upon the percent cumulative variance in the PCA analysis. We consulted with a domain expert to verify that the data patterns observed from PCA were also observable in the laboratory. Our results were confirmed by observational data, and that indeed these were observed times of high and low activity.

Results for PCA can be seen in Figures 2 and Figure 3. Figure 2 shows a three dimensional vector representation of the locations of each animal for window 3. The axes are labeled with the three principal components that were found to be significant. The dots show the location of the window data points in the PCA projected space. Each line indicates specific animals of interest, such as the breeding male (B), animal 17, 3, 12, etc. The center collection of vectors shows those animals that share similar behavior patterns in all three components. The same three dimensional graph for window 14, figure 3, shows much more activity. This showed that animal behavior in windows 3 and 14 were different. Whereas behavior patterns in window 3 could be explained by fewer behaviors (e.g. sleeping, moving about, and going to the toilet cage), behavior patterns in window 14 were much more complex and included other behaviors (perhaps digging, and searching for food).

Typically, behaviorists use central graph analysis to show interactions between pairs of animals. The central graph is a representative graph based on a set of graphs in which an edge is created if it is present in 50% or more of the graph set. (Banks and Carley 1994) In the face of this huge body of data, central graph analysis was not tractable using the entire data set. Thus in order to do central graph analysis it was necessary to sample the dataset. In previous work we were able to show that frequent pattern mining yielded similar results to central graph. Efficiencies due to the pruning done by frequent pattern algorithms enabled data analysis using the entire dataset. (McCloskey et. al. 2011). It was shown that frequent pattern mining was useful in showing social interactions between not only pairs of NMRs, but larger groups as well by creating sociograms based on the results of frequent pattern analysis.

Association rule mining has been used in market basket analysis for finding groups of items that are purchased together. If we think of our sensors as corresponding to the cash registers in a supermarket, it is easy to map the association rule market basket problem to the problem of finding socializing groups of NMRs. Each store transaction corresponds to groups of items in a customer's market basket. If we look at many market baskets, we find that certain sets of items tend to be frequently purchased together. From these we can find rules to the effect, "customers that buy shampoo are likely to buy soap". These rules hold with minimum statistical significance and confidence. Our goal for this study was to find mole rats that group together, given the sensor location data. For this analysis we were interested in the composition of groups, not group location. To transform the data for use by the Apriori algorithm, we partitioned each observation in the state matrix by sensor locations. Each individual partition can be viewed as a "transaction" that occurred at a sensor. If we look at many of these transactions we can find 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_n} 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. If an itemset satisfies some minimum percentage of

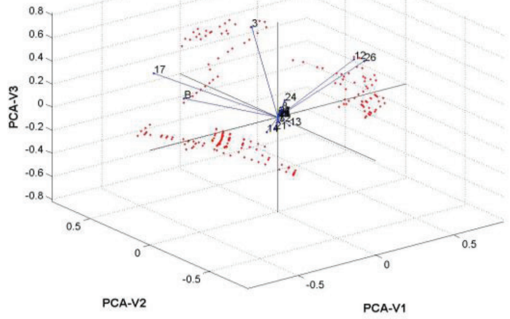

Figure 2 Three dimensional graph of data projected onto PCA components for window 3

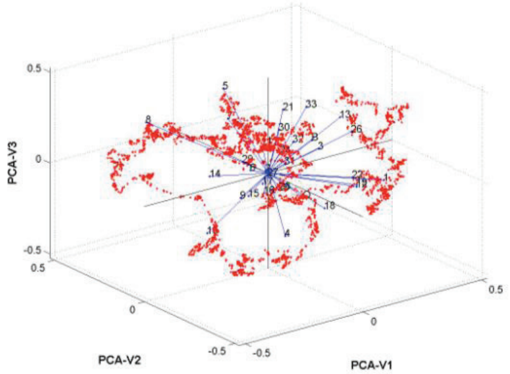

Figure 3 Three dimensional graph of data projected onto PCA components for window 14

transactions it is called a frequent itemset. The percentage value is known as the support of the itemset. The absolute support of an itemset is the actual number of transactions satisfied by that itemset in the dataset. Support is an indication of the itemset's statistical significance (Agrawal, Imielinski, and Swami 1993) (Agrawal and Srikant 1994) (Agrawal et. al. 1996).

obs#	NMR1	NMR2	NMR3	NMR4	NMR5
1	1	2	1	2	3
2	1	2	1	3	3

Table 1 Sample state matrix for 5NMR, 2 observations, and 3 sensors

obs#	NMR1	NMR2	NMR3	NMR4	NMR5
1	1	0	1	0	0
1	0	1	0	1	0
1	0	0	0	0	1
2	1	0	1	0	0
2	0	1	0	0	0
2	0	0	0	1	1

Table 2 Transformed state matrix for frequent pattern mining.

Since our "transaction" data represented a partitioning of each row in the state matrix, any one NMR could be seen in one out of fourteen transactions, given a fourteen sensor configuration. Given an observational dataset, and a set of items I, where each observation records a value for each item in I, the values for each observation partitions the set of items, I, such that for each observation we have subsets $X_1, X_2, X_3, \ldots, X_n$ where n is the number of subsets, and $X_1 \cap X_2 \cap X_3 \cap \ldots \cap X_n = \phi$, and $X_1 \cup X_2 \cup X_3 \cup \ldots \cup X_n = I$. Define absolute partitioned support as the number of times the itemset satisfies a transaction divided by n, the number of partitioned subsets. Define partitioned support as percent defined by the absolute support divided by the number of observations in the original state matrix. Partitioned support is the statistical significance of an itemset relative to the original state matrix. Table 1 shows a sample state matrix for five NMR. Table 2 shows how the data was transformed to be used with the Apriori algorithm.

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 (Pasquier, N. et. al. 1999), (Wang, Han, and Pei, 2003) and maximal itemsets (Gouda, K., and Zaki,, M.J. 2001). Given a frequent itemset X, X is a closed frequent itemset if there exists no $Y \supset X$ 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. (Gouda, K., and Zaki,, M.J. 2001) 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, NMR5 ⇒ NMR24 has no different confidence measure than NMR24 ⇒ NMR5 since both NMR24, and NMR5 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. (Zaki 2000) One of the advantages of frequent pattern analysis over central graph is the ability to look at varying levels of support, or statistical significance. Central graph uses a 50% cutoff for its consensus matrix. In analyzing data from window 3 and window 14, we were able to find itemsets of frequent social groupings, at different support levels. We ran our analysis at partitioned support levels of 20%, 50%, and 80%. This approach effectively allows us to determine what proportion of time any dyad of animals spends together during a time window, and ultimately determines the strength of every relationship.

Social psychologists use a graphical model called a sociogram to show interactions between agents in a given population. This form of analysis is powerful since it allows for secondary measures of the strength of connected nodes. (Butts, 2008), (Freeman, 1979) (Wasserman, and Faust, 1994). Klemettinen et. al. describe a method of visualizing association rules using a graphical representation. Each attribute is represented by a node, and directed arcs between these nodes represent rules. Building on this, we can create similar graphs using maximal itemsets. Since the maximal itemsets capture groups of NMRs that socialize, we can thus use an undirected graph to visualize these itemsets. We represent each animal as a node in the graph and edges connect NMRs that are in the same maximal frequent 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.

Conclusions

Figures 5-7 show sociograms derived from frequent pattern analysis with 20%, 50%, and 80% support in window 3 while figures 8-10 show the social interactions at different levels of support in window 14. The sociograms were created using the graphical software package, Gephi (http://gephi.org). The sociograms confirm the results obtained from PCA. Windows 3 and 14 show markedly different numbers of interactions between animals.

Overall, there are more social interactions during window 3, than during window 14, when the animal behavior patterns were more variable. Additionally, although fewer interactions occur in window 3, as one increases support, over all, the number of dyad interactions

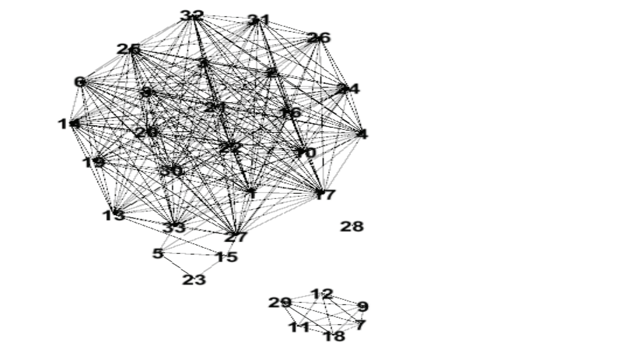

Figure 5 Sociogram of Window 3 with 20% support

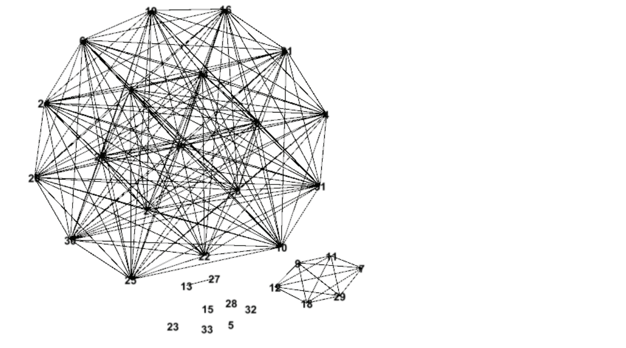

Figure 6 Sociogram of Window 3 with 50% support

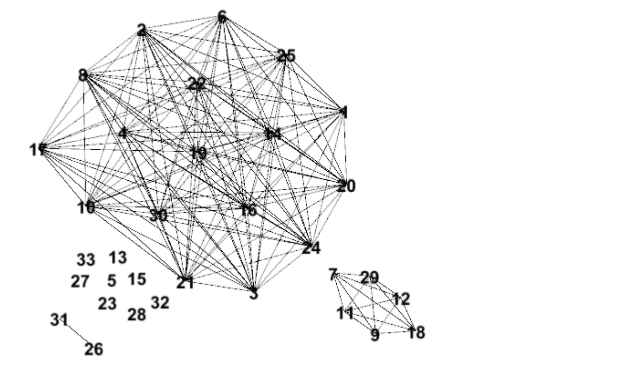

Figure7 Sociogram of Window 3 with 80% support

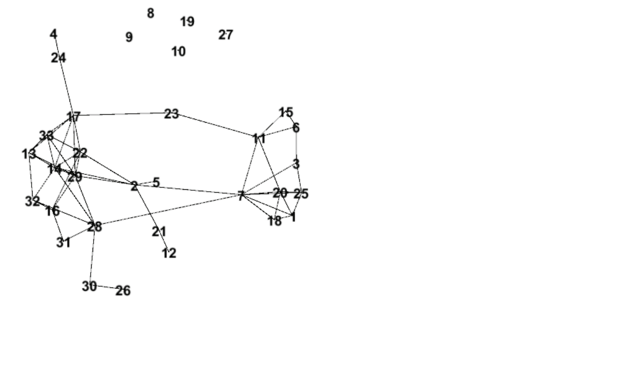

Figure 8 Sociogram of Window 14 with 20% support

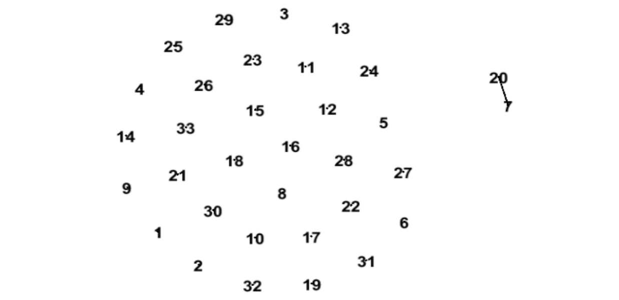

Figure 9 Sociogram of Window 14 with 50% support

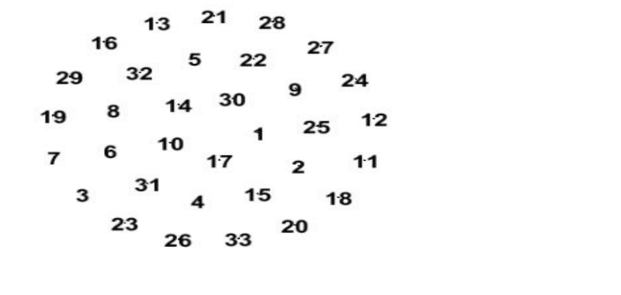

Figure 10 Sociogram of Window 14 with 80% support

does not decrease significantly from the 20% support level to the 80% support level. One interesting observation is that animals that socialize infrequently at 20% tend to have no interactions at higher support levels. This might indicate that the number of social interactions may be correlated with the frequency of social interaction. Thus non social NMR not only interact with fewer NMR, they interact less frequently than more interactive NMR. In addition, comparison of the associations in window 3 and window 14 provide evidence that supports our hypothesis that social relationships are determined by the type of activity pattern. For example, animal 28 does not show any social interactions at any level of support during window 3, but does exhibit social behavior when support is set to 20% during window 14. In fact, animal 28 provides the link between animals 30 and 26 to the rest of the group during this time, exhibiting a quality known as betweeness centrality (Wasserman and Faust, 1994). Conversely, animals 20 and 7 show the only dyad at the 50% support level in window 14, but do not associate at any level of support during window 3. These variations in social behavior may provide clues to the roles of individuals in colony maintenance and protection. These results show, for the first time that social behavior is dynamic and dependent upon the activity of the larger group..

Summary

In this paper, we have shown that a combination of PCA and frequent pattern mining can be used to demonstrate both the variability of behavior in large groups and the interactions among members of that group simultaneously. An important advance of this paper is that this method allowed for the depiction of frequent itemsets as a sociogram that depicted dyad relationships with varying levels of statistical significance, or support. Our data is the result of using RFID technology to observe animals in their housing environment. This technology generated huge amounts of data that could not be analyzed using the traditional methods employed by social psychologists. Our results indicate that frequent pattern mining offers social psychologists a new way of analyzing animal behavior data.

For the future, we intend to analyze data over multiple days to identify groups of animals that socialize on a daily basis. We continue to record data from the colony and intend to analyze this data looking for weekly, monthly, and seasonal patterns. We intend to use techniques employed in social network analysis to look more closely at the sociograms. By tagging other items in the environment such as food and pups, we can get insights into the hierarchical behavior, such as the specific roles for soldiers and workers. In short the techniques described here have opened up a way to observe laboratory animals with a precision that has not yet been done before.

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 wish to thank Marion Felix on her work on the PCA analysis. Thanks to Julien Cerard, Michael Cicero, and Michael Vazquez for their work.

References

Agrawal, R., and Srikant, R. 1994. Fast Algorithms for Mining Association Rules. IBM Res. Rep. RJ9839. IBM Almaden.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A. I. 1996. Fast Discovery Of Association Rules. In U. Fayyad, G. Piatetsky Shapiro, P. Smyth, & R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining (pp. s 307 328). AAAI/MIT Press.

Banks, D.L., and Carley, K.M. (1994). ``Metric Inference for Social Networks.'' Journal of Classification, 11(1), 121 49.

Brett, R. A. 1991. The population structure of naked mole rat colonies. In P. W. Sherman, J. U. M. Jarvis & R. D. Alexander (Eds.), The Biology of the Naked Mole Rat (pp. 97 136) Princeton, NJ: Princeton University Press.

Buffenstein, R. 2008.. Negligible senescence in the longest living rodent, the naked mole rat: Insights from a successfully aging species. Journal of Comparative Physiology.B, Biochemical, Systemic, and Environmental Physiology, 178(4), 439 445.

Butts, C.T., 2008. Social network analysis: A methodological introduction. Asian Journal of Social Psycholog. 11, 13 41.

Clarke, F.M., and Faulkes, C.G., 1999. Kin discrimination and female mate choice in the naked mole rat Heterocephalus glaber. *Proceedings of the Royal Society of London B*; 266 (1432), 1995 2002.

Freeman, L.C., 1979. Centrality in social networks: Conceptual clarification. Social Networks 1(3), 223 258.

Gouda, K., and Zaki,, M.J. 2001. Efficiently mining maximal frequent itemsets. *ICDM, First IEEE International Conference on Data Mining* (p. 163).

Klemettinen, M. Mannila, H., Ronkainen, P., Toivonen, H., and A. Verkamo, I., 1994. Finding interesting rules from large sets of discovered association rules. In *Proceedings of the third international conference on Information and knowledge management* (CIKM '94), Nabil R. Adam, Bharat K. Bhargava, and Yelena Yesha (Eds.). ACM, New York, NY, USA, 401 407.

Kritzler, M., Lewejohann, L., and Krüger, A. 2007. Analysing movement and behavioural patterns of laboratory mice in a semi natural environment based on data collected via RFID technology. *Proceedings of the Workshop on Behaviour Monitoring and Interpretation*: 17 28

Kritzler, M., Jabs, S., Kegel, P., and Krüger, A. 2008. Indoor tracking of laboratory mice via an RFID tracking framework. In *Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS less environments* (MELT '08). ACM, New York, NY, USA,

Lewejohann, L., Hoppmann, A.M., Kegel, P., Kritzler, M., Kruger, A., and Sachser, N. 2009. Behavioral phenotyping of a murine model of alzheimer's disease in a seminaturalistic environment using RFID tracking. Behavioral Research Methods 41:850 856.

Martinez, Wendy L., and Martinez Angel R., *Computational Statistics Handbook with MATLAB*, 2nd Edition 2002. Boca Raton: Chapman & Hall/CRC

McCloskey, D. P., M. E., Imberman, S. P., Kushnir, I, and Briffa Mirabella, S.. 2011. From market baskets to mole rats: using data mining techniques to analyze RFID data describing laboratory animal behavior. in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '11). ACM, New York, NY, USA, 301 306.

O'Riain, M.J., Jarvis, J.U., Alexander, R., Buffenstein, R., and Peters, C. 2000. Morphological castes in a vertebrate. *Proceedings of the National Academy of Sciences of the United States of America*, 97(24), 13194 13197

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. 1999. Discovering Frequent Closed Itemsets for Association Rules. In *Proceedings of the 7th international Conference on Database theory*. In C. Beeri & P. Buneman (Eds.), (LNCS 1540, pp. 398 416).

Wang, J. , Han, J., and Pei, J.. 2003. CLOSET+: searching for the best strategies for mining frequent closed itemsets. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '03). ACM, New York, NY, USA, 236 245

Wasserman, S., and Faust, K. 1994. *Social network analysis: Methods and applications*. Cambridge; New York: Cambridge University Press.

Zaki, M. J. 2000. Generating non redundant association rules. In *Proceedings of the Sixth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (Boston, Massachusetts, United States, August 20 23, 2000).