Undergraduate Research: Membership in the Knowledge Discovery in Databases (KDD) Research Group has provided the opportunity to be involved in several major research projects. I initially became interested in link mining research with an emphasis on the link structure of social networks. In order to study the link structure of a social network I first created a web crawler *LJCrawler* that quickly collected user and inter-user information from the social network service *LiveJournal*. By studying the results, I found that this data could accurately be represented as a directed graph. Next I created *LJMiner* which was able to extract and construct features about candidate user-pairs. Three types of features were considered: user independent, pair-dependent, and graph features. I then constructed a learning algorithm which learned a classifier based on these features. This classifier accurately predicted the existence of annotated links that were intentionally removed. As I experimented with various configurations of learnable features I found that the inclusion and omission of certain features resulted in the construction of vastly different classifiers – a finding shared by Dr. Robert Holte in his 2007 presentation at ECML. As a result of this work, I co-authored a paper[1] describing our scientific methodology and results that was accepted into AAAI's 2006 Spring Symposium. As an extension, I co-authored separate papers[2,3] which used an evolutionary approach to select features for link mining.

Further advances in this area led me to select the link mining topic as my undergraduate honors research project. Specifically, my goal was to develop foundational theories for deriving causal relationships on a large scale. To that end, I revamped *LJCrawler* and *LJMiner* to operate on very large graphs and I developed optimized graph algorithms to compute inter-user and graph features from within the large graph. As a result of this work, I co-authored[4] and presented a paper describing my work at the 2007 ICWSM. A third iteration of this work studied the temporal phenomenon of these graphs. I authored[5] and presented a paper describing the temporal methodologies and results, and was the session chair, at the 2008 ANNIE Conference.

These papers and presentations began for me an exciting adventure in research. I began to receive correspondence asking for implementation details and more information regarding my published work, and I have happily assisted other researchers from differing countries in reproducing my results on varying domains.

Graduate Research: During the summer of 2008 I was selected to attend the Discrete Sciences Summer Institute's Multimodal Information Access and Synthesis (MIAS) Center at the University of Illinois Urbana-Champaign. MIAS is concerned with researching technologies for extracting and tracking interesting events. I led a team of graduate and undergraduate students in developing a search engine we affectionately called *Webster*. Specifically, *Webster* employed state of the art clustering and machine learning techniques to do information extraction and synthesis of UIUC sports' web pages. I developed several textual and semantic analysis modules that were enhanced by my link mining expertise; the output from these modules were entities that, when analyzed against each other, could be assimilated into a master entity-relation (ER) model. Finally, the ER model could be searched or browsed with a web browser.

The most common problem I encountered was caused by noisy data from crawled web pages because of the abundance of advertisements and menus. I needed a way to extract only the content of a web page. I solved this problem by developing an algorithm that looks at the structure of the HTML source code to cluster and extract content. Not only was this beneficial to the project, but I also authored[6] and presented a paper describing my findings at the 2008 DEXA

Conference in Turin, Italy. After the conclusion of the summer I was able to broaden the scope of the content extraction algorithm to the more general case of clustering 1-dimensional spaces. I again authored[7] and presented a paper describing my findings.

More recently, I have worked on exploring inductive biases that exist when performing link mining. Part of that exploration has led to collaboration with association rule researchers. Together, we formulated a new approach that considers the relative number of items (e.g., links, interests, groceries) and appropriately adjusts the recommendation's confidence. I co-authored[8] and presented a paper at the 2008 ANNIE Conference detailing the algorithm and results.

In the large graphs that we study, algorithms, such as Dijkstra's shortest path algorithm, perform very poorly, yet my link mining algorithms are required to perform several shortest path calculations. To avoid the expensive costs of conventional algorithms I devised an approximation technique which considers low-dimensional embeddings of the graph. Preliminarily, we show an improvement from Dijkstra's quadratic-time algorithm to an accurate logarithmic-time approximation. Publications are pending at ACM's KDD Conference.

The common thread of all my past research is that relationships between entities are not always explicitly indicated, and that link mining algorithms can provide an insight into relationships that are inherently present in many domains. Furthermore, I believe this experience has prepared me well for my proposed research into feature discovery in relational domains.

Publications:

[1] Hsu W. H., King A. L., Paradesi M. S. R., Pydimarri T., Weninger T. "Collaborative And Structural Recommendation Of Friends Using Weblog-Based Social Network Analysis", Proc. of Computational Approaches to Analyzing Weblogs - AAAI 2006 Technical Report SS-06-03, pp. 55-60, Stanford, CA, March 2006.
[2] Hsu W. H., King A. L., Paradesi M. S. R., Pydimarri T., Weninger T. "Evolutionary Data Mining For Link Analysis: Preliminary Experiments On A Social Network Test Bed", GECCO-2006. Seattle, WA, July 2006.
[3] Hsu W. H., Lancaster J., Paradesi M. S. R., & Weninger T. "Collaborative and Structural Recommendation of Friends using Weblog-Based Social Network Analysis", GECCO-2007. London, July 7-11, 2007.
[4] Hsu W. H., Lancaster J., Paradesi M. S. R., & Weninger T. "Structural Link Analysis from User Profiles and Friends Networks: A Feature Construction Approach", ICWSM-2007. Boulder, CO, March 26-28, 2007.
[5] Weninger T., Hsu W. H., Paradesi, M. S. R. "Predicting Links and Link Change in Friends Networks - Supervised Time Series Learning with Imbalanced Data", ANNIE-2008. St. Louis, MO, Nov 9-12, 2008.
[6] Weninger T., Hsu W. H. "Text Extraction from the Web via Text-To-Tag Ratio", DEXA-2008 Workshop on Text-based Information Retrieval. Turin, Italy, Sept 1-5, 2008.
[7] Weninger T., Hsu W. H. "Web Content Extraction Through Histogram Clustering", ANNIE-2008. St. Louis, MO, Nov 9-12, 2008.
[8] Al-Jandal W., Weninger T., Hsu W. H. "Validation-Based Normalization and Selection of Interestingness Measures for Association Rules", ANNIE-2008. St. Louis, MO, Nov 9-12, 2008.

Presentations were given for publications 4, 5, 6 and 7. Two more papers are under review.