Predictive analytics that takes in account network relations: A case study of research data of a contemporary university

Ekta Nankani[1] **Simeon Simoff**[1,2]

[1] School of Computing & Mathematics
University of Western Sydney,
Email: enankani@scm.uws.edu.au

[2] Email: s.simoff@uws.edu.au

Abstract

Contemporary organisations incorporate large amount of invisible networks between their employees. The structure of such networks impacts the information fusion within the organisation. Taking into account the influence of such network structures in predictive modeling will be beneficial for the quality of organisational strategic planning. Network mining methods (the social network analysis of large heterogeneous data sets) can extract information about the structure of such networks and the strategic positioning of each individual from various interaction data. We propose to integrate the output of network mining into the predictive modeling cycle in order to depict these influences. This paper demonstrates such approach by incorporating network centrality measures of actor closeness and actor betweeness in CART predictive modeling cycle. It presents a proof-of-concept application of this integrated approach to the case study of a contemporary university, which resembles some similarity with corporate organisations. The study utilises a data set about academic research activities collected over five years. The results of the study support the hypothesis that information about the network structures in a data set (whose impact is included through the centrality measures) can improve the accuracy of predictive analysis.

Keywords: Predictive Analytics, Social Network Analysis (SNA), Centrality Measures, Data Enrichment

1 Introduction

The research direction taken in this work has been inspired by the visionary research by Tapscott and Williams (Tapscott & Williams 2006) on challenging the deeply rooted assumptions about the role of competitiveness and collaboration in business and society as a whole. "The four principles – openness, peering, sharing and acting globally – increasingly define how twenty-first-century corporations compete." ((Tapscott & Williams 2006), p.30). Their new economic vision draws a picture of a world of a business collaboration on a massive scale as a key to survive in a globally competitive environment. Remaining innovative requires understanding the shifts in the environment and the development of new strategies that foster collaboration in order to progress in a competitive environment. This reality faces both industry and academia. Technology advances are based on the advance of fundamental sciences. Contemporary research and development activities in industry are tighten to the need of being fast, efficient and capable of earning clear return on investment. However, innovations continue to rely on fundamental knowledge, hence, industry will increasingly rely on partnerships with universities and other research organisations, leaving corporate research teams to move quickly to technology development and practical application. In practice, close cooperation between industry and academia potentially can enable the industry partners to keep their edge, while spreading the upfront research and development costs across a much broader ecosystem (see (Tennenhouse 2004) for an example of the implementation of such strategies).

An essential strategy in making the most out of such partnerships is the deepening and broadening collaboration across research communities, starting with fostering strategic collaborations within a university. This is the practical problem that motivates the research in predictive modeling presented in this paper.

Understanding the structure of existing and predicting potentially new collaborations is vital when it comes to enabling the interaction between industry and academia. This interaction as well the interaction and combination of several disciplines, are seen as the key drivers of contemporary innovation. Consequently, critical becomes the development of robust business intelligence methods that can

- extract essential information and knowledge about the structure of collaboration;
- produce reliable models that can be used for prediction (recommendation) of new collaborative ventures.

In depicting collaboration these analytics methods and respective technologies have to deal with heterogeneous data about academic activities that link academics into various invisible networks.

In this paper we have focused on predictive modeling that contributes information to the processes that support the development of research directions in universities. The paper presents early proof-of-concept results in two aspects of academic activity:

1. obtaining internal research funding, and;
2. type of research output in terms of publication categories.

The first one looks at predicting whether a research project proposal, submitted to one of the university research grant schemas will be funded or not. The second one looks at predicting the most-likely DEST category of publications in which academic publications will fall into, e.g whether these publications will be books, book chapters, journal articles or conference papers.

Both of these tasks involve researchers from the same organisation, hence, the assumption is that *the structure of collaborative relations between researchers matters.* Consequently, the paper explores two ways of predictive modeling for each of the above tasks:

- conventional predictive modeling without taking the structure of collaborative relations;
- extended predictive modeling, which takes in account information about the structure of relations.

The presentation in the paper is centred around this practical problem in order to demonstrate the practical value of proposed solutions. Further the paper is organised as follows: Section 2 looks at two modeling perspectives - predictive analytics and social network analysis, in the context of the problem. Section 3 considers some of network centrality measures as carriers of information about the positioning of and relations between network structural elements. Section 4 uses a dataset about research activities of academics in an Australian university to present the integrated approach and methodology for addressing the above formulated predictive tasks; Section 4.5 discusses the results of the analysis; and Section5 considers the future developments and concludes the paper.

2 Modeling perspectives in analytics

Contemporary knowledge economies and digital ecosystems rely on capturing and utilising diverse data about the activities and processes within them. Consequently, a continuously growing variety of data analytics techniques addresses the need for converting these data into useful information for decision making purposes. This section provides a brief overview of the two modeling perspectives in analytics that are relevant to presented work: predictive modeling (predictive analytics) and social network analysis for competitive intelligence.

2.1 Modeling perspective in predictive analytics

In data analysis models which are used to predict future data trends are known as predictive analysis models. Classification or estimation algorithms are central in predictive analytics and are used in many areas of human endeavour, including (but not limited to) business and science. Examples of application areas from business include credit approval, medical diagnosis, performance prediction and selective marketing. Predictive models assess unlabeled samples to determine the value or value ranges of an attribute that a sample is likely to have(Han & Kamber 2001). With predictive analysis the validity of the classification result (and the true accuracy of the model) can be verified by waiting for the future event to happen. Though predictive accuracy is a critical aspect of models there are other facets that are equally important. We may require a model to show which of the predictor variables are most important in the dataset(Smyth 2001). We may be interested in examining whether predictor variables interact or whether a simple model can result in good prediction. In the research, presented in this paper, we are interested in taking in account the structure of “social” relationships between the entities in a predictive modeling dataset. In particular, we consider enriching the predictive modeling dataset with attributes that represent information about the structure of such relationships. Such attributes are based on concepts from social network analysis (SNA). In this paper we append attributes that correspond to some SNA centrality measures and then test the hypothesis that *appending centrality measures improves the prediction accuracy.* For the purpose of the paper, the dataset we use is a snapshot of a five year span, that, to some extent, encapsulates the temporal relationship of predictors to the target variable(Linoff 2004).

Any of the classification or estimation techniques can be used for predictive analysis on the proposed enriched dataset. The five criteria for evaluating predictive methods include predictive accuracy, computational speed, robustness, scalability and interpretability(Han & Kamber 2001). Proposed enrichment of the dataset affects four of these criteria. On the positive side is the expected improvement of predictive accuracy and interpretability of the results. However, the approach requires additional computation of centrality measures, which will affect the computational speed and scalability.

We show in Figure 1 an example of a fragment of data about academics and research students in a university, similar to the one considered in the case study, to position the approach presented in the paper. The data set includes the following attributes: Name (of the person), Position (in the university), School (as administrative unit), Research Center (for those involved in research centers), Publication type (according to DEST classification), Co-Authorship (indicates a list of people from the same set that have published together with the person in consideration), Co-Supervision (indicates a list of people from the same set that have co-supervised higher degree research students).

‘Conventional’ predictive modeling deals with the portion of the data set, contoured by the double dotted line. Let the task be the prediction of the type of publication in which most off the output of a researcher will be falling into. Hence, the attribute “Publication type” is selected as the “output” (“target”) attribute and the attributes “Position”, “School”, “Research Center” are the “input” (“predictors”). As the aim is to derive a general trend, the attributes that contain unique identifiers, such as “Name”, will not be taken in account. As a result, ‘conventional’ predictive modeling cycle does not have mechanisms to take into account some of the relations that may exist between the instances in the data set - in our case, between the individuals. The issues and problems of depicting such dependencies with predictive analytics methods have been discussed in the context of network mining in (Simoff & Galloway 2008). The chapter considered two groups of issues:

1. the ”loss of detail” - the hidden links existing between the instances in a data set;
2. the assumption about the independency of the attributes of a data set.

In this paper we deal with networks that are explicitly encoded. For instance, a co-authorship relation between two researchers can be define as the association of the names as authors on the same paper. Though it may not accurately and in-depth reflect the actual authorship in terms of contribution and development of the research work, it reflects the underlying assumption that co-authorship involves some interaction and information exchange between the authors. In terms of data analysis this means some embedded dependency between the instances that represent these researchers in the data set. Centrality measures are one way to represent explicitly this dependency. Next section presents aspects of social network analysis relevant to the approach presented in the paper.

2.2 Modeling perspective in social network analysis

Social networks represent groups of people, various connections among them and the dynamics of such connections for in-depth analysis (McDonald April 2003). The production of knowledge is a social process involving interactions among people and organisations with different backgrounds, resources, predispositions and insights(von Krogh et al. 2001, Tushman & Rosenkopf 1992). Measuring these heterogeneous social networks is done to study the influence of emergent social structures within and external to an organisation on the business and engineering processes within it.

Social network analysis and network mining are means that address the problem of discovering organisational intelligence from existing and potential interactions in the organisational settings. Traditional social network analysis usually deals with networks where only "cognitive agents" (people, groups of people, the human capital of organisations) can be the nodes. Network mining can be viewed as an extension of SNA, not just in terms of the volume of the data, but also in terms of the content of the network models: nodes can be any elements, including resources, expertise, intellectual property, technologies, products, markets).

According to Wasserman and Faust, "Social Network Analysis (SNA) provides a formal conceptual means for thinking about social world" (Wasserman & Faust 1994). Contemporary SNA deals with the analysis of interactions between social entities in an organisation, based on large data sets of human interactions (Shetty & Adibi 2005). SNA research recognises the elements of an organisation as intentional networks, hidden networks, socially translucent networks, mediators, and structural holes. These elements can depict changes with processes in a group over a period of time. Important for our work is that through such elements and the respective model parameters in network analysis methods we obtain information about the structural inter-dependence in an organisation, that is, "who knows who", and "who knows what" (Srivastava et al. 2006), which to some extent reveals structures of information fusion. This comes from the fact that our daily life is very much influenced by social networks through which we interact with various groups of people: family, friends, colleagues. Through these networks we indirectly connect to people associated with these groups without necessary knowing them. The need to take into account these interactions has been recognised in several areas of applied modeling in information systems, including viral marketing, e-mail filtering based on social networks, various recommender systems (Matsuo et al. 2007).

The study of social networks formed on social networking sites, such as orkut, flickr, youtube, myspace, can help to detect the most influential users. Many properties of the social network have been studied: Pool and Kochen scientifically formulated the small world phenomena(Schnettler 2009); according to Milgram the average path between two Americans is six hops (Schnettler 2009); Granovetter suggest that social networks can be partitioned into strong and weak ties, with strong ties tightly clustered (Granovetter 1973); nodes with high indegree also tend to have high outdegree, showing active members are also popular members (Mislove et al. 2007).

Study of social influence is a strategic arena for SNA research. Some argue that influence is a special instance of causality, namely the variations of one person's responses by the actions of another(Stanley Wasserman 1994). SNA approach and techniques are not limited to humans and can be used to study a variety of phenomena (Wasserman & Faust 1994), hence the increased interest in the academic community (Kumar et al. 2006). Contemporary SNA is associated mostly with visual analysis of graph structures(McDonald April 2003).

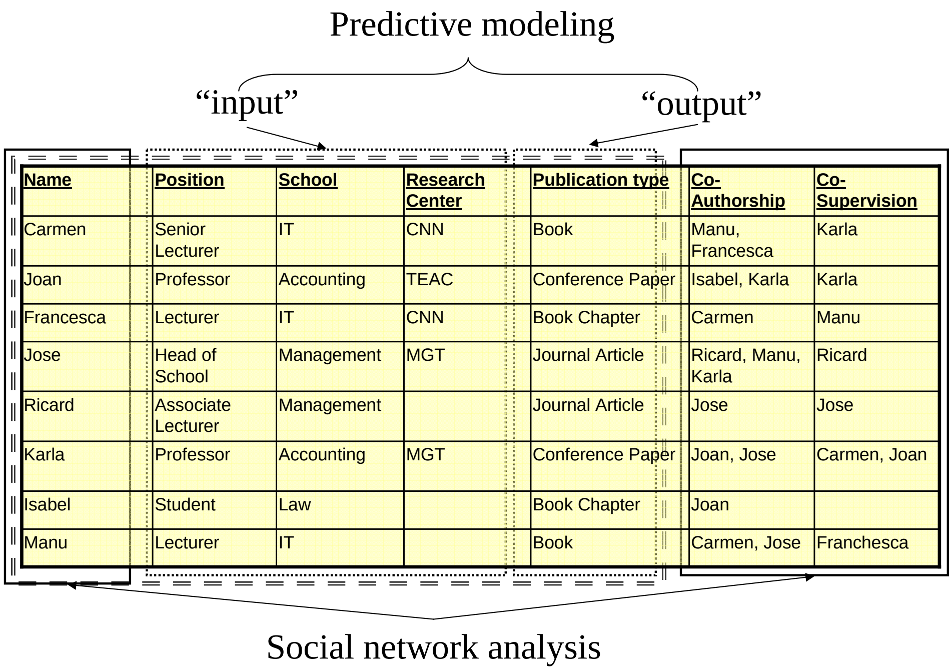
As mentioned in Section 2.1 our method brings the SNA modeling perspective into predictive modeling. It considers the estimated underlying graph models from a portion of a data set that usually would be ignored in predictive modeling cycle. Several parameters of such models are included in the extended data set for predictive modeling. The practical grounds for taking such approach are motivated from previous network studies which indicate that the social structural contexts surrounding actors shape a variety of responses both attitudinal and behavioural. In customer analytics, for example, behavioural features are believed to be more reliable than demographics. Behavioural targeting is to target right person at right time, hence the drive for developing methods that can produce more accurate predictions of customer behaviour. Logically such methods should utilise information from various networks in which customers can be involved, including alumni, referrals, rehires and business development (Drakos et al. 2008). In an organisation the inferred networks assist in identifying the knowledge flow and find out solutions for corporate related problems, sometimes even the extent to which an individual has succeeded in performing his work (Heer 2004).

In this paper we consider the utilisation of information about collaborative networks, which are important drivers of the knowledge flows within organisations (Singh 2005), including universities and research institutions. Scientific networks are an example of collaborative networks that has a long history of investigation, in particular, citation networks have been studied as knowledge flow structures in sciento- and bibliometrics. More recently, the focus has shifted to co-authorship networks in order to get a better understanding of the underlying structure of knowledge evolution. Relevant to the underlying philosophy of our work is the use of a regression method to estimate the probability of knowledge flow between inventors of any two patents (Hu & Jaffeb 2003).

Our work is also inspired by "the law of the few" or the "80/20 principle" (Gladwell 2000). According to Gladwell, "the success of any kind of social epidemic is heavily dependent on the involvement of people with a particular and rare set of social gifts." In other words in any situation roughly 80 percent of the 'work' will be done by 20 percent of the participants. Gladwell divides these '20 percent' into three types:

- *Connectors* - those ones that "bringing the world together" as a result of their ability to span many different worlds;
- *Mavens* - those ones that connect people with new information, i.e. the information brokers;
- *Salesmen* - those ones that persuade people.

The difference between these types of actors in social networks is reflected in their positions and patterns of linking. Hence, the inclusion of social network measures in the training dataset enables the capture of such information in the predictive model. In the next section we discuss some of the centrality measures, that have been utilised in this study. Rather



Name	Position	School	Research Center	Publication type	Co-Authorship	Co-Supervision
Carmen	Senior Lecturer	IT	CNN	Book	Manu, Francesca	Karla
Joan	Professor	Accounting	TEAC	Conference Paper	Isabel, Karla	Karla
Francesca	Lecturer	IT	CNN	Book Chapter	Carmen	Manu
Jose	Head of School	Management	MGT	Journal Article	Ricard, Manu, Karla	Ricard
Ricard	Associate Lecturer	Management		Journal Article	Jose	Jose
Karla	Professor	Accounting	MGT	Conference Paper	Joan, Jose	Carmen, Joan
Isabel	Student	Law		Book Chapter	Joan	
Manu	Lecturer	IT		Book	Carmen, Jose	Franchesca

Figure 1: “Predictive modeling” and “social network analysis” perspectives of a data set

than from a graph-theoretical point of view, we discuss the role of these measures in the network models from a domain perspective.

3 Information about network structural elements reflecting relations

When dealing with publications in broad sense, including not only papers but also postings on various usenet groups and blog sites, the analysis of network relations is as essential as the text analysis of the content. For instance, in (Agrawal et al. 2003) researchers demonstrated that link analysis can be more valuable than text-based algorithms when it comes to classification of people on two sides of an issue in a usenet group.

Analysis of centrality measures determines the importance of vertices in a network based on their connectivity within the network structures. For instance, in health science centrality measures help researchers in depicting the structure of underlying biological networks that model biological processes as complex systems; the approach has been successfully applied to different biological networks (Dwyer et al. 2006).

Social network relations are measured within a set of actors. In this paper we consider a single mode network - a network described by a dataset that contains information about only one type of actors - in this case these are people. The actors include academic staff, researchers, students and externals, associated with university. The relationship between actors is a kind of professional collaboration, and includes *co-authorship*, *co-supervision*, *co-teaching*, and *co-participation* in a project.

There are different measures to quantify network relationships. These measures help to test propositions about network properties rather than simply relying on descriptive statements. To understand the role of an actor in a network SNA evaluates the location of actors (nodes) through a set of centrality measures. These measures provide information about the different aspects of actors' role in a network with respect to their position, e.g. connectors, bridges, leaders, isolates, as well as about the clusters in the network structure and which actors are in them, which actors form the core of the network, and which actors reside on the the periphery.

Centrality of an actor is measured in terms of actor *degree*, *closeness* and *betweeness*. *Actor degree* refers to the number of links an actor has. The idea behind actor degree is that actors with more links are in a more independent position - such actors are less dependent on any specific actor. In terms of collaborative research networks, high values of actor degree measures may indicate more administrative research role (e.g. a research director) than a research collaborator role in terms of ideas flow, hence actor degree measures are not taken in account in the current work.

Actor closeness measures the ability of an actor to reach other actors in a network at shorter path lengths, or, reciprocally, actors who are more reachable by other actors at shorter path lengths. In terms of collaborative research networks this structural advantage can be translated into potential for initiating research collaboration, e.g. starting a project or initiating co-supervisory arrangements.

Actor betweeness measures the ability of an actor to broker contacts among other actors in the network, e.g. the extent to which an actor is positioned between the other actors. In terms of collaborative research networks this structural advantage can be translated into potential for growing research collaboration, e.g. extending an existing team of chief investigators for the next grant application, amalgamating research groups into a larger entity.

In this study we consider four centrality measures:

- three closeness measures: closeness, eigenvector centrality and harmonic closeness, and;
- betweeness.

The brief description of these measures is presented below following (Wasserman & Faust 1994).

Closeness measured as the length of the shortest-path, scores higher values to more central vertices. Closeness at actor n_i level is calculated as

$$C_C(n_i) = [\sum_{j=1}^{g} d(n_i, n_j)]^{-1}$$

where, $C_C(n_i)$ is *Closeness* of n_i and $d(n_i, n_j)$ is the number of links in the geodesic path linking actors i and j (that is $d(node1, node2)$ is a distance function) and this sum is from $j = 1$ to $j = g$, where g refers to all the other actors not including i actor. This index is the inverse of the sum of the distances from actor i to all other actors. In terms of information flow, those actors with highest closeness values are well positioned for monitoring the information flow in the network. In a collaborative research network research leaders are expected to be in such positions.

Eigenvector centrality (known also as eigenvector of geodesic distances) is another form of closeness, looking for the most central actors in terms of the overall structure of the network. From a factor analysis perspective, the eigenvector centrality measure ranks actors in terms of some new dimensions that characterise the distances among actors, where the first of this new dimensions captures the positioning of an actor with respect to the overall network structure, and the rest are depicting more local substructures. An eigenvalue in this context defines the location of each actor with respect to each dimension, hence, the term eigenvector when considered with respect to all actors in the network. The measure of centrality is computed as the largest positive eigenvalue. The eigenvector centrality measure for n_i is

$$C_{EV}(n_i) = \sum (C_{EVmax} - C(n_j))/C_{EVmax}$$

where, $C_{EV}(n_i)$ is eigen vector for n_i, C_{EVmax} is max eigen vector and this sum is for all the actors from $i = 1$ to $i = j$.

Harmonic closeness is an alternative measure of closeness that takes in account all the pathways that connect an actor to all others, rather than just the geodesic. The measure estimation is based on an algorithm that uses the harmonic mean length of paths ending at the given node. In a collaborative research networks broad collaborators are expected to be in positions with high value of this measure.

Betweenness depicts those actors that occur on many shortest paths between other actors, having higher betweenness than those that do not. Similar to the other centrality measures, there is a family of betweenness measures - the one used in our study is

$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i)/g_{jk}$$

where, $C_B(n_i)$ is Betweenness of n_i and g_{jk} is geodesic linking two actors i and j. The actor betweeness centrality for n_i is sum of estimated probabilities over all pairs of actors not including ith actor. Actors with high betweeness can be power players, but can be also the single point of failure. In a collaborative research network, for instance, their removal may cause fragmenting (up to disintegration) of the network.

In an earlier work (Nankani, Simoff, Denize & Young 2009) we have focused on the discovery and analysis of network structures in university data about academic activities. The method relied on a combination of network mining techniques with substantial visual analysis and qualitative data analysis for validation purposes. The work has analysed networks at different levels of granularity, varying from individual level through to networks between divisions. In this paper we use only the network structures of relations between individuals. In the next section we use a case study format to test the hypothesis that information about the network structures in a data set (whose impact is included through the centrality measures) can improve the accuracy of predictive analysis.

4 Case study of university research data set

The case study in this section is based on an integrated university research data. It demonstrates the integrated approach of social network mining combined with predictive analysis on two predictive modeling tasks:

1. *forecast internal research grant application outcome* - whether a research project will get funding or not, and;
2. *predict the predominant category of personal publication output* - whether an academic will be publishing predominantly conference papers, journal articles, book chapters, books or any other category of creative work registered in the data set.

The completion of both tasks depends on numerous factors beyond the scope of the dataset. By incorporating the centrality measures we are looking at developing a feasible approximation for performing these predictions.

4.1 Description of the data set

Table 1 shows the description of the data set, which includes integrated data about a range of different academic activities, including

- co-authorship;
- co-participation in a research project;
- co-supervision of research students;
- other related academic data, which is not taken in account in this work.

All data are time-stamped, collected over a consecutive span of 5 years. During this period of time the university in consideration has had 9 schools and 23 research centres. All collaborative ties are between staff, students and external participants.

Readers can find more details about some of the results of the network analysis in (Nankani, Simoff, Young & Denize 2009) [these include details about the evolution of the networks over a time span and analysis of centrality measures, with network visualisations generated with NetDraw graph visualisation tool].

4.2 Methodology

We divided this project into three different stages, as shown in Figure 2

The purpose of each of these phases is detailed as follows.

Phase 1 includes integrated data collection, cleaning, developing an understanding of the data structures and composing the original data set for the analytics tasks. Details of Phase 1 are discussed in (Young et al. 2008, Nankani, Simoff, Denize & Young 2009, Nankani, Simoff, Young & Denize 2009).

Phase 2 includes

Data Statistics

Description	Number of Records
Data Records collected	24,556
Clean Data Records	15,177
Number of Distinct Nodes	2,131
Number of Ties	37,398

Table 1: Description of the university research data set

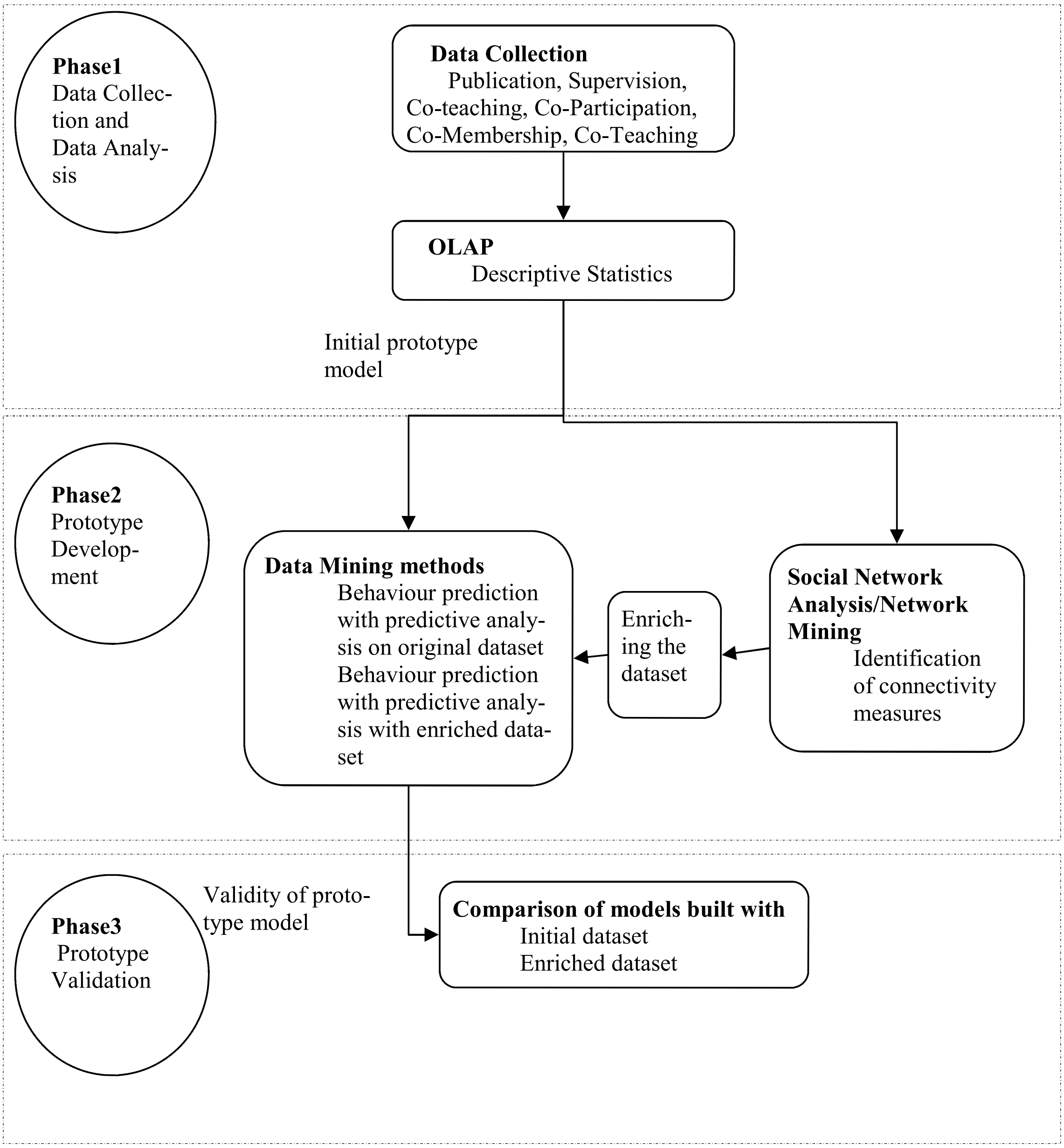

Phase1
Data Collection and Data Analysis
Data Collection
Publication, Supervision, Co-teaching, Co-Participation, Co-Membership, Co-Teaching
OLAP
Descriptive Statistics
Initial prototype model
Phase2
Prototype Development
Data Mining methods
Behaviour prediction with predictive analysis on original dataset
Behaviour prediction with predictive analysis with enriched dataset
Enriching the dataset
Social Network Analysis/Network Mining
Identification of connectivity measures
Phase3
Prototype Validation
Validity of prototype model
Comparison of models built with
Initial dataset
Enriched dataset

Figure 2: Methodology

1. creating predictive model based on the the original data set;
2. performing social network analysis based on the original dataset (work presented in in (Young et al. 2008, Nankani, Simoff, Denize & Young 2009, Nankani, Simoff, Young & Denize 2009);
3. enhancing the original data set by appending centrality measures;
4. creating predictive models based on the enhanced dataset.

Phase 3 includes the analysis and comparison of the models, created with the original and enhanced data sets.

4.3 Generating centrality measures

The centrality network measures of closeness, eigenvector, harmonic closeness and betweeness are estimated with the respective algorithm implementations in UCINET social network analysis software. These network measures then are appended as additional attributes to the existing academic data set, extending the dataset with data about the network structures.

4.4 Predictive modeling

For this study we looked at type of classifiers that have relatively poor predictive power, but are good in handling mixed types of data and missing values, and are insensitive to monotone transformation of inputs and robust to outliers in the input space. Last but not least - classifiers with good level of interpretability of the results. Based on these criteria we have selected tree classifiers as in general they have poor predictive power and meet the rest of the criteria (see (Hastie et al. 2001), Table 10.1). CART classification tool by Salford Systems(Salford_Systems n.d.) is well suited for the purpose of the study, as it implements classification and regression trees (for some details see (Linoff 2004)). Figure 3 illustrates the steps taken to create the predictive models that address tasks 1 and 2 discussed in the beginning of section 4.

4.4.1 Predicting project funding (Task 1)

These models involved 13 attributes from the original data set, including `Person Name`; `Person Code`; `Type` (with the following possible values: 'internal member' (from the same university), 'external member' (from other university or industry) and 'student'); `Faculty` (to which the person belong to), `School` (to which the person belong to); `Year` (when a project started or a publication was made); `Research center membership`; `Project name`; `Project Status` (whether funding grant application has been approved or rejected); `Publication category`.

Model for predicting project funding based on the original data set

Predictive models were created with a sample from the original data set (sample of records are taken because of restriction of software used). All the project data was divided into 4 files with 800 records in each data set. Several random sets of 800 records with exclusion were taken at a time and respective predictive models were created and tested. The model created with the following attributes to predict Project Status whether project will be approved (funded) or rejected (not funded) scored the best result. The attributes used with the model include `Research center membership`; `Person Name`, `School Name`, `Project Name`; `Year` and `Faculty Name`.

Variable Importance

Importance	Original dataset	Enhanced dataset
1	Person Name	ProjectName
2	Project Name	Research Center
3	School Name	Local Eigenvector
4	Faculty Name	Closeness
5	Year	Harmonic Closeness
6	Research Center	Betweenness

Table 2: Varible Importance - Project Funding

Predictive Accuracy
Learn Dataset - 81.46%
Test Data Set Using
Cross Validation - 54.62%
Separate test file - 55.77%

Model for predicting project funding based on the enhanced data set

Another set of predictive analysis models are created with enhanced data set (sample of records are taken because of software restriction). A random 800 records were taken and several predictive models were created. The model created with the following attributes to predict `Project Status` whether project will be approved (funded) or rejected (not funded) scored the best result. The attributes used with the model are `Project Name`, `Closeness`, `Betweeness`, `Eigen Vector`, `Harmonic closeness`, `Local Eigenvector`, `Year`, `Research center membership`.

Predictive Accuracy
Learn Dataset - 64.83%
Test Data Set Using
Cross Validation - 60.32%
Test data - 60.72%

Variable importance for predicting project funded or not funded

Table 2 displays all the variables in order of importance, with most important as 1 to least important as 6, for the analytical models created with the help of project data set.

4.4.2 Predicting publication category (Task 2)

This task started with the same data set as for Task 1.

Model for predicting publication category code based on the original data set

Predictive analysis models for publication category code are created with the publication data set

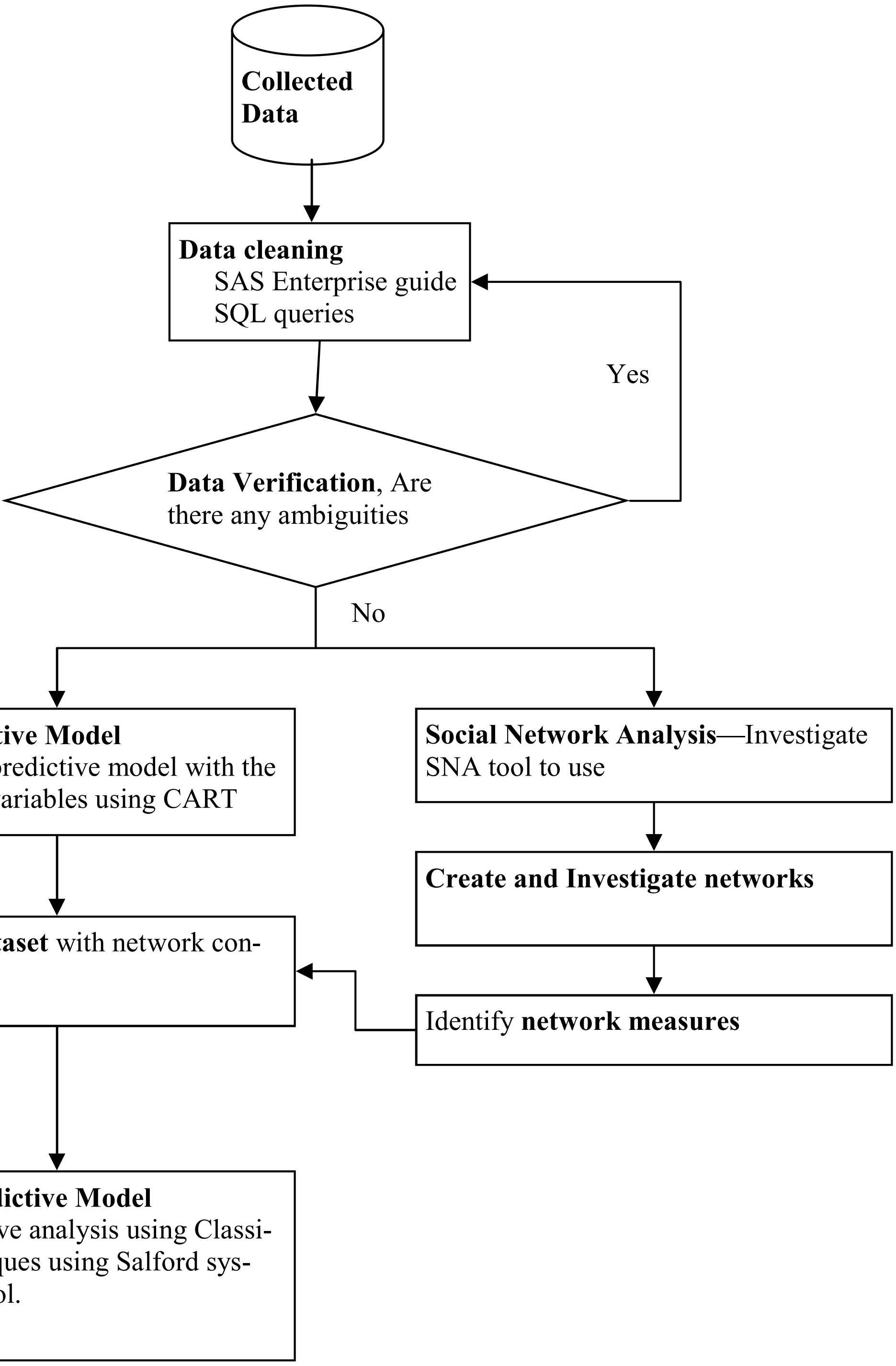

Collected Data
Data cleaning
SAS Enterprise guide
SQL queries
Yes
Data Verification, Are there any ambiguities
No
Initial Predictive Model
Create predictive model with the help of initial variables using CART
Social Network Analysis—Investigate SNA tool to use
Create and Investigate networks
Enrich the dataset with network connections
Identify network measures
Enriched Predictive Model
Predictive analysis using Classification techniques using Salford systems CART tool.

Figure 3: Predictive Modeling Steps

Variable Importance

Importance	Original dataset	Enhanced dataset
1	Person Name	Closeness
2	School Name	Harmonic Closeness
3	Faculty Name	Faculty
4	Year	Betweeness
5	Research Center	Year
6	Type	Eigenvector

Table 3: Varible Importance - Publication Category

from the original data and a sample of records are taken (sample of records are taken because of software restriction). A random 800 records were taken in a dataset and several predictive models were created and tested. All publication datset is divided into 10 data files. Models are tested in several iterations to include all records for model testing purpose. The model created with the following attributes to predict publication category code scored the best result. The attributes used with the model include `Research center membership`; `Person Name`, `School Name`, `Year`, `Faculty Name`, `Type` (Person type).

Predictive Accuracy
Learn Dataset - 83.69%
Test Data Set Using
Cross Validation - 68.42%
Separate test file - 69.72%

Model for predicting publication category code based on the enhanced data set

Another predictive analysis models are created with enhanced data set (sample of records are taken because of software restriction). Again a random 800 records were taken and several predictive models were created and tested. The model created with the following attributes to predict publication category scored the best result. The attributes used with the model include `Closeness`, `Betweeness`, `Eigenvector`, `Harmonic closeness`, `Local Eigenvector`, `Faculty Name`

Predictive Accuracy
Learn Dataset - 78.6%
Test Data Set Using
Cross Validation - 76.53%
Test data - 74.10%

Variable importance for predicting publication category code

Table 3 displays all the variables in order of importance, with most important as 1 to least important as 6, for the analytical models created with the help of project data set.

4.5 Comparison of the results

A summary of the results from the analytical models in this pilot study are presented in Table 4. The results from the experiments with the original and extended data sets in Task 1 are presented in columns *Task 1.o* and *Task 1.e.* The results from the experiments with the original and extended data sets in Task 2 are presented in columns *Task 2.o* and *Task 2.e.* These results illustrate that when the estimated SNA centrality measures of one part of the data set are added as complementary predictors to the other part of the data set they improve the prediction accuracy both in a cross validation setting and on unseen data.

Therefore, this preliminary study supports the hypothesis that information about the network structures in a data set can improve the accuracy of predictive analysis. To some extent this approach can be viewed as enhanced predictive analytics technique.

5 Conclusions

Since the days of the six-degree separation experiment, social network analysis has advanced significantly, thanks to the prevalence of online social websites and their capabilities of collecting data about the communities created around them, as well as the availability of a variety of offline large-scale social network systems such as collaboration networks. There are several technologies to support rich social interactions as blogs, wikis, social bookmarks, social tagging and these techniques are finding their way into business environments (Drakos et al. 2008).

This paper addressed the issue of utilising the information about the network structures of relations between the instances of a data set in predictive modeling cycle. Such practical problems emerge in various corporate settings, as well as in academic collaboration in universities.

The work presented a method that deploys SNA methods for extracting the structure of the network. Essential information about this structure is encoded through the various network centrality measures. In this work we have depicted four measures.

The results of this study support the hypothesis that information about the network structures in a data set (whose impact is included through the centrality measures) can improve the accuracy of predictive analysis. In both predictive tasks we have improved the average percentage accuracy over the test data. Though the improvement in the accuracy is several percent, the additional data preprocessing for estimating respective centrality measures is worth considering in domains like cancer treatment, where every percent of increased accuracy matters!

References

Agrawal, R., Rajagopalan, S., Srikant, R. & Xu, Y. (2003), Mining newsgroups using networks arising from social behavior, *in* 'WWW '03: Proceedings of the 12th international conference on World Wide Web', ACM, New York, NY, USA, pp. 529–535.

Drakos, N., Mann, J., Cain, M. W., Andrews, W., Knox, R. E., Valdes, R., Rozwell, C., Bradley, A., Maoz, M., Otter, T., Harris, K., McGuire, M., Bell, T., Basso, M., Prentice, B., Smith, D. M., Fenn, J., Prentice, S., Sarner, A., Dunne, M. & Harris, M. (2008), Hype cycle for social software, Research ID Number: G00158239, Gartner. The Social Software Hype Cycle highlights the most important technologies that support rich social interactions. Use our assessment of their business relevance and maturity to guide your investment decisions.

Dwyer, T., Hong, S.-H., Koschützki, D., Schreiber, F. & Xu, K. (2006), Visual analysis of network centralities, *in* 'APVis '06: Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation', Australian Computer Society, Inc., Darlinghurst, Australia, Australia, pp. 189–197.

Datasets used for prediction	**Predictive accuracy (%)**			
	Task 1.o	**Task 1.e**	**Task 2.o**	**Task 2.e**
Learn dataset	81.46	64.83	83.69	78.6
Test dataset using cross validation	54.62	60.32	68.42	76.53
Test dataset using test data	55.77	60.72	69.72	74.10

Table 4: Summary of the results

Gladwell, M. (2000), *The tipping point: How little things can make a big difference*, Little Brown.

Granovetter, M. S. (1973), 'The strength of weak ties', *The American Journal of Sociology* **78**(6), 1360–1380.

Han, J. & Kamber, M. (2001), *Data Mining: Concepts and Techniques*, Morgan Kaufmann.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The elements of statistical learning*, Springer.

Heer, J. . (2004), Exploring enron: Visualizing anlp results, *in* 'University of California, Berkely'.

Hu, A. G. Z. & Jaffeb, A. B. (2003), 'Patent citations and international knowledge flow: the cases of korea and taiwan', *International Journal of Industrial Organization* **21**(6), 849–880.

Kumar, R., Novak, J. & Tomkins, A. (2006), Structure and evolution of online social networks, *in* 'KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, New York, NY, USA, pp. 611–617.

Linoff, M. J. A. B. G. (2004), *Data Mining Techniques: for marketing, sales, and customer relationship management*, Wiley Publishing.

Matsuo, Y., Mori, J., Hamasaki, M., Nishimura, T., Takeda, H., Hasida, K. & Ishizuka, M. (2007), 'Polyphonet: An advanced social network extraction system from the web', *Web Semant.* **5**(4), 262–278.

McDonald, D. (April 2003), 'Recmmending collaboration with social networks: A comparative evaluation', *ACM* **5**, 5–10.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P. & Bhattacharjee, B. (2007), Measurement and analysis of online social networks, *in* 'IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement', ACM, New York, NY, USA, pp. 29–42.

Nankani, E., Simoff, S., Denize, S. & Young, L. (2009), Enterprise university as a digital ecosystem: Visual analysis of academic collaboration, *in* 'DEST2009'.

Nankani, E., Simoff, S., Young, L. & Denize, S. (2009), *Information Systems: Modeling, Development, and Integration*, Springer Berlin Heidelberg, chapter Supporting Strategic Decision Making in an Enterprise University Through Detecting Patterns of Academic Collaboration, pp. 496–507.

Salford_Systems (n.d.), Cart a robust decision tree tool for data mining, predictive modelling, data preprocessing, Technical report, Salford Systems.
URL: *www.salfordsystems.com/doc/CARTtrifold.pdf*

Schnettler, S. (2009), 'A structured overview of 50 years of small-world research', *Social Networks* **31**(3), 165–178.

Shetty, J. & Adibi, J. (2005), Discovering important nodes through graph entropy the case of enron email database, *in* 'ACM, KDD 2005'. Chicago.

Simoff, S. J. & Galloway, J. (2008), Visual discovery of network patterns of interaction between attributes, *in* S. J. Simoff, M. H. Boehlen & A. Mazeika, eds, 'Visual Data Mining: Theory, Techniques and Tools for Visual Analytics', Vol. 4404 of *LNCS*, Springer Verlag, Heidelberg., pp. 172–195.

Singh, J. (2005), 'Collaborative networks as determinants of knowledge diffusion patterns', *Management Science* **51**(5), 756–770.
URL: *http://ssrn.com/paper=628281*

Smyth, D. H. . H. M. . P. (2001), *Principles of Data Mining*, MIT.

Srivastava, J., Pathak, N., Mane, S. & Contractor, N. S. (2006), Knowledge perception analysis in a social network, *in* 'Workshop on link analysis counter terrorism and security 22nd Apr 2006, Bethesda, Maryland.'.

Stanley Wasserman, J. G., ed. (1994), *Advances in Social Network Analysis: Research in the social and behavioral sciences*, SAGE.

Tapscott, D. & Williams, A. D. (2006), *Wikinomics: How Mass Collaboration Changes Everything*, Portfolio, Penguin Group.

Tennenhouse, D. (2004), 'Intel's open collaboration model industry-university partnerships', *Research Technology Management* **47**(4).

Tushman, M. & Rosenkopf, L. (1992), *Organizational Determinants of Technological Change: Toward a sociology of Technological Evolution*, Vol. 14, Greenwich CT: JAI Press.

von Krogh, G., Nonaka, I. & Aben, M. (2001), 'Making the most of your company's knowledge: A strategic framework', *Long Range Planning* **34**(4), 421 – 439.

Wasserman, S. & Faust, K. (1994), *Social Network Analysis: Methods and Applications*, cambridge University Press.

Young, L., Simoff, S., Denize, S. & Nankani, E. (2008), Who collaborates with whom and why? exploring the typography and evolution of collaborative networks. IMP Conference 2008, "Studies on business interaction? Consequences for business in theory and business in practice".