

Chapter XIV

A Case Study in Data Mining for Automated Building of Teams

Robert K. McCormack

Aptima, Inc., USA

Andrew Duchon

Aptima, Inc., USA

Alexandra Geyer

Aptima, Inc., USA

Kara Orvis

Aptima, Inc., USA

ABSTRACT

This chapter highlights a case study involving research into the science of building teams. Accomplishment of mission goals requires team members to not only possess the required technical skills but also the ability to collaborate effectively. The authors describe a research project that aims to develop an automated staffing system. Any such system requires a large amount of personal information about the potential team members under consideration. Gathering, storing, and applying this data raises a spectrum of concerns, from social and ethical implications, to technical hurdles. The authors hope to highlight these concerns by focusing on their research efforts which include obtaining and using employee data within a small business.

INTRODUCTION

Data mining of customer information is now widespread throughout the business world, however systematic data mining of employee information is less common. Thus, while the opinions of consumers have driven much of the effort to protect their data, the opinions of employees have often been ignored. The attitude of employees towards the use of their personal data is greatly affected by their perceptions of the company's intentions. If the employer makes the effort to obtain permission from individual employees and carefully explain the goals of the data collection, negative perceptions may be reduced (Long & Troutt, 2003). Even when legal restrictions are observed, data mining within one's own organization can have far-reaching social and ethical implications. Any process which can be considered invasive to personal data requires sensitivity to all stakeholders, including both the employees and the company itself (Saban, 2001).

This chapter highlights the approach taken by one company in gathering, storing, and using employee data to automatically staff teams for new projects. The company described herein is a privately owned research and development firm with approximately 100 employees. The majority of the staff is comprised of scientific researchers with graduate degrees in psychology, cognitive science, human-system engineering, modeling and simulation, and computer science.

Much of the business performed by the company is supported by dozens of small government contracts obtained by responding to quarterly requests for proposals (RFPs) and winning follow-on work. Currently, teams for writing the proposals and carrying out the project work are created by word of mouth. This chapter focuses on one such project, TeamBuilder, conducting research into the science of building teams automatically from employee data. The goal of the project is to make the process of staffing teams more efficient and the projects more successful

by automatically finding people that both have the requisite skills and would work best together as a team. It was decided by the TeamBuilder research team to use their own company as a test bed, due to familiarity with team processes within the organization and availability of data. It was determined early in the project that, in order to establish whether each given candidate possesses the necessary skills and abilities that are required for a specific team, large amounts of personal data would be needed.

Using employee data for any purpose has both drawbacks and benefits. Clearly, the employer would like to use all available information about employees that would make them more efficient, productive and successful. A well-conducted data analyses could also help build better relationships between employees and the organization, increase the opportunities and choices available to individuals, and in general help build a better understanding of the organization as a whole. Additionally, automated interpretation of the data could help management reach conclusions beyond the ability of human analysis and avoid inefficient word-of-mouth processes. On the other hand, there is the ever-present concern about invasion of privacy. Exploiting personal information also brings the risks of false conclusions being drawn, abuse of information to the detriment of individuals or organizations, and misapplication of erroneous data. These risks must be weighed against the benefits (Cook & Cook, 2003).

The organization of the chapter is as follows. We will first discuss the scientific and theoretical issues of staffing teams that are the basis for the TeamBuilder project. Then we will describe how we gathered the data for our case study while addressing privacy, ethical, legal and security concerns of the organization (both employees and management). Because the goal of the project is to create teams automatically, the integrity of the data must be high in order to obtain the trust of both management for using the tool and employees for understanding their assignments. We therefore

describe a number of processes undertaken to ensure data integrity. Finally, we briefly discuss some of the computational methods we use to operationally define team theoretical constructs and employee abilities, in order to automatically staff teams with people who both have the skills to do the task and will work well together. There are issues with applying any automated process to assess human behavior and we provide some details of how we mitigate these concerns.

BACKGROUND

TeamBuilder Overview

Teams are an integral part of almost every organization. The use of teams allows consideration of expertise in multiple areas (Rouse, Cannon-Bowers, & Salas, 1992) as team members are often brought together with diverse knowledge, skills, and abilities. With technological advancements in communication, teams can now be formed across barriers of time and space, as project requirements dictate. The greater wealth of information that exists in teams, in comparison to any one individual member, has prompted a plethora of research dedicated to better understand the nature of teams (see Kozlowski & Ilgen, 2006, for a review). The area of team research that is very relevant to the TeamBuilder project is concerned with the understanding of team composition.

Traditionally teams have been staffed by matching individual demographic characteristics (training, rank, experience) to generic functional roles and known project requirements (Klimoski & Jones, 1995; Klimoski & Zukin, 1999). However, little research or practical attention has been directed to evaluating the efficacy of particular staffing strategies or determining how well team members, selected under certain strategies, actually work together as a team on given projects and within specific mission parameters. Ideally, team composition should reflect the full range of

performance requirements posed by both the task itself and the collaborative quality of teamwork. **Taskwork skills** are the capabilities necessary to effectively complete all of the performance requirements for a particular task or project. **Teamwork skills**, on the other hand, are capabilities that allow individuals to operate effectively in a multi-person environment. Teamwork skills are the mix of attitudes, personality, and values that would optimize teamwork effectiveness and group cohesion among a particular set of individuals working within particular contexts. (Cannon-Bowers, Tannenbaum, Salas, & Volpe, 1995; Klimoski & Zukin, 1999; Morgan & Lasiter, 1992; Salas, Burke, & Cannon-Bowers, 2002).

A team staffing strategy should also reflect the most effective means of matching available human capital with the full range of requisite taskwork and teamwork skills demanded by the team's mission. While traditional staffing strategies have considered taskwork requirements, few, if any, have similarly acknowledged teamwork needs (Klimoski & Jones, 1995; Klimoski & Zukin, 1999). For example, if a project required a lot of interaction between the team members, then one might want to staff that project with people who have the generic teamwork skill of *cooperation* (Kerr, 1983).

The purpose of this research and development effort is to create a team staffing tool which assists managers in automatically forming teams that fulfill both taskwork and teamwork requirements. One of the sub-goals of this project is to capitalize on existing organizational resources (e.g. resumes, project information, and employee communication logs) to help identify the individuals who possess the skills required by project. A major aspect of this effort then is to determine the computational means by which to link the raw data to abstract taskwork and teamwork skills. Additionally, we discuss the ethical, legal, privacy and social implications that arose as various data resources were mined and applied in this context.

GATHERING STORING AND USING DATA FOR STAFFING TEAMS

Gathering the Data

Before the data-gathering process began, we first needed to address several issues regarding the ethical, legal, privacy, and social implications of collecting and using employee data. Although the company owns many of the data sources we wanted to utilize for TeamBuilder, such as time-card records, email, instant messaging, and phone logs, it was quickly decided that participation in the project would be strictly on an “opt-in” basis. With the support of the executive management, and assistance from legal professionals, a consent form was given to each employee describing the purpose of the project, their rights, and the types of data being collected. While many employees had concerns about privacy and the creation of an Orwellian atmosphere, explanation of the purpose and processes of the project allayed most of these fears. In the end, an overwhelming majority of employees chose to participate and allow access to their data. In this section, we will discuss the data gathering process for this project, including ethical and legal hurdles, the types of data collected, and the social implications of gathering data.

Addressing Privacy, Ethical, and Legal Risks of Data Gathering

Prior to gathering data, an effort was made to identify the risks and stakeholders in order to mitigate future problems. The data that had to be implemented in TeamBuilder could be divided into two categories: information about employees and information about projects. In general, gathering and using employee and project information introduces a variety of concerns. Whereas the exposure of project data poses little risk to individual employees, it could prove detrimental to the company itself. Use of employee data, on

the other hand, raises many issues, including privacy concerns for both the individual and the company as a whole. In order to alleviate some of the privacy issues, it was decided not to collect personal information regarding age, gender, race, performance, or pay because they were not essential for the purpose of this effort. Only employee and project data deemed absolutely necessary for TeamBuilder would be collected.

In order to create a version of TeamBuilder which would aid company staffers in putting together the best teams for proposals, projects, and special committees, it was necessary to obtain and analyze several data resources. Among these resources, resumes and biographies provide textual descriptions of individual professional skills, educational background, and employment history. This type of information is vital when it comes to matching the needs of the mission to the capabilities of the employees. Resources such as e-mail, phone, and instant messaging logs provide insight into the communication network among company employees. Rather than affording skill-based assessments of each individual, the communication logs can be used to determine teamwork skills. Timecards provide historic snapshots of project teams, the level of involvement by various employees, and how much groups of employees have worked together which also touches on teamwork skills.

Project data supplies detailed information regarding all the past projects and teams that have been formed in the company. Spreadsheets on projects and proposals feature titles of the efforts, names of managers, budget information, period of performance, and other important facts. Descriptions of project requirements are obtained from government RFPs (Requests for Proposals). These requests describe the technical needs, dates, and provide names and contact information of customers for the given project. RFPs are used in TeamBuilder as a basis for matching project requirements to individual skills. Table 1 illustrates the data resources utilized in the

Table 1. Summary of data resources

	Data Resource	Description	Format
Employee Information	Resumes	Resumes for current and past employees with information on skills, past employment, and education	Unstructured text; format varies widely from person to person
	Biographies	Short description of employees' current and past work	Unstructured text
	Hire dates	Dates of hire and termination (if applicable) for current and past employees	Structured spreadsheet
	Communication logs	E-mail, phone, and instant messaging (IM) logs with information on sender, recipient, date/time, and size/duration (content not gathered)	SQL database
Project Information	Timecards	Hours charged by each employee to various projects broken down month by month	Structured spreadsheet
	Project information	Summary information on each project, including title, dates, customer, budget, and manager	Structured spreadsheet
	Proposal information	Summary information on each proposal, including title, dates, customer, and proposal leader	Structured spreadsheet
	Request for proposals (RFP)	Government issued requests with descriptions of desired work	Unstructured text

initial development of TeamBuilder, as well as their individual description.

Taking under consideration the social and ethical concerns that might be raised by gathering above data resources, efforts were made ensure that the collection was handled with utmost sensitivity and respect for privacy. It has been shown that individual's perceptions of fairness and invasion of privacy are affected by his or her ability to authorize or disallow disclosure of information (Eddy, Stone, & Stone-Romero, 1999). Even though all data required for the TeamBuilder project is the legal property of the company (since it is maintained on company computer equipment, systems and/or networks and used for business purposes), the company decided to extend a professional courtesy to its employees by allowing them to decide if they

wanted the data related to them to be used in this study. As Cook and Cook (2003) point out, ethics are more restrictive than the law and adhering to the law does not always mean being ethical. Even though the company was not legally obligated to obtain consent forms from its employees, it was done regardless.

The consent form was handed out to all company employees and contained relevant information about the project (e.g. funding source, purpose, benefits) as well as reasons for needing to collect the data. Several of the possible risks were identified, and the steps that would be taken to prevent them were described. It was also noted that participants could at any time during the course of the study revoke their consent, and withdraw without prejudice, and that refusing to participate would involve no penalty or loss of

benefits to which they were entitled as employees. If a participant chose to withdraw from the experiment, personal data from that participant would be removed from all databases.

The following points illustrate some of the risks and concerns, as well as the mitigation strategies, as they were presented in the consent form.

1. Your signing of the informed consent form does NOT allow us to use the data for any other application/reason than building and testing the validity of our TeamBuilder approach.
2. If TeamBuilder proves to be a useful tool, it will NOT be the ultimate decision maker for all teams created. It is intended to be a decision support tool that can help people choose members for teams (whether that would be proposal teams, special interest teams, project teams, etc.). The decision to use it would be strictly up to you. Currently we build teams by word of mouth and employee expertise is sometimes overlooked. Our hope is that it will open more opportunities up to company employees, and not less. In addition, functionality is built into the tool which allows novices to try new roles/projects. Therefore, TeamBuilder will not stunt employee growth.
3. We are not gathering any content from E-mail, IM, and phone. Even if we wanted to look at the content of those messages, we couldn't feasibly do it. We have absolutely no access to it whatsoever. We are using the "to-from" data for social network analysis, to see who is communicating with whom. This may serve as selection criteria if the "mission" required a team that could hit the ground running. In that case, you might want to pull together people who have worked together many times in the past.
4. Your data will not be shared with anyone outside of the organization. We are creating "dummy data" for the customer.

This approach was successful in that about 93% of the company's employees signed the consent form granting us access to their data. Of those who did not participate, some stated that they made this decision because they felt uncomfortable sharing their private data, even if it was only for the purpose of the TeamBuilder project. Overall, the key concerns for many employees were with respect to accessing the content of their e-mail, IM, and/or phone conversations. We reassured people that this would not be the case. We also stressed to everyone that refusal to sign the consent form would not result in any type of penalty or loss of benefits.

For several members of the staff, it was deemed in the best interest of the individuals, the company, and the TeamBuilder project not to include them in the data collection process. These employees included several members of the human resources staff who regularly send and receive highly sensitive emails and communications. Even though we were not gathering content of these communications, the frequency or amount of emails between HR staff and individual employees may itself be highly sensitive and raise privacy concerns. As Fule and Roddick (2004) explain, data at different levels of sensitivity can be represented in a hierarchy of interest. The frequency with which email is sent may be less sensitive than the full content of the messages, but these levels of sensitivity can vary among different groups.

Gathering Project and Employee Data

With the help of the human resource staff, contracts management, and information systems groups, gathering the various data resources was a straightforward procedure. Based on the list of identified resources as well as the list of participating and excluded employees, data resources were collected and stored in a central computer. Since employee biographies (displayed on the company website) and government RFPs are publicly available, they were easily obtained.

Project and proposal information, while containing somewhat sensitive company data, is available to all employees. With the approval of the executive management, these resources were obtained for the TeamBuilder project. Resumes, hire dates, timecards, and communication logs were collected and filtered to exclude the non-participating employees.

Storing and Organizing the Data

After obtaining consent and collecting the various data sources, the next step in the TeamBuilder project was to incorporate the data into a single relational database. In this section we will discuss the technical challenges in transferring data from raw formats into a cohesive database structure while preserving the integrity of the data. This includes structured data such as email logs and timecards, as well as unstructured text, as found in resumes and project abstracts. We will delve into the process of correlating disparate data sources in which identifiers do not always match (e.g. names spelled differently in separate data sources). In addition, we will discuss the steps taken to ensure privacy of the data.

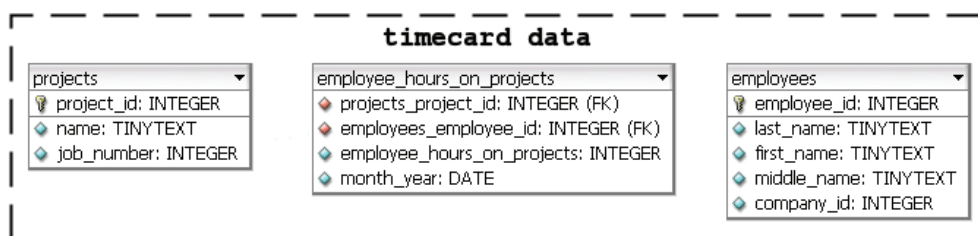
In all, we identified and gathered eight different types of data, as described in Table 1. Because each data source varied widely in the type and format of information contained within, individual data extractors were created to parse

out the relevant information from the source and store it in a relational database table. Note that this is inherently an ad hoc process. Extractors were tailored for the individual idiosyncrasies of each data source and relied on us recognizing the repeating patterns, if any. Once the desired data were identified in the text, they were inserted into the proper fields of the database.

Figure 1 displays a sample of the database tables derived from timecard data used for this project. Each box represents a table in the TeamBuilder database. The column names and data types (e.g., INTEGER vs. TEXT) are defined for each table. The data types used for each table were often dictated by the original data source.

The timecard data is broken out into three separate tables: employees, projects, and a relational table describing the hours each employee worked on a given project during a particular month. Because the timecard management system uses automated record keeping, it was decided that this data would serve as the canonical list of both employees and projects. That is, the timecard data is taken as the standardized list of all employees and projects to which other data resources are mapped. The correlation of disparate data resources is discussed shortly. Now, we briefly describe how each data resource was mined, and the types of information extracted.

Figure 1. Example tables in the TeamBuilder database



Parsing and Storing Unstructured Textual Data

Much of the data needed for TeamBuilder is stored in unstructured documents. These include resumes, biographies, and government RFPs. Each of these types poses unique challenges for extracting data. Because they consist mainly of free text and have little consistency in format between documents, locating relevant information can be time consuming. Here we describe the general approach to data extraction from each unstructured data source.

- *Resumes*: Employee resumes vary widely in their formats, but generally contain the same types of information. For this project, we were interested in the skills and past education of each employee. In the analysis and use of resume data, text analytic techniques are used (as described in the next section). Therefore, it is not necessary to manually identify and categorize different features of the text, such as specific skills, because the text analytic algorithms operate on raw text. However, sections of the resumes which describe skills, education, etc. must be tagged and stored in the database for inclusion in the algorithms. Because the resumes lacked a cohesive format, much of data tagging was done by hand.
- *Biographies*: Employee biographies are short, one to two paragraph descriptions of current and past projects and skills, used mainly for marketing (e.g. placement on the public website or attaching to proposals). Like the resumes, biographies are used to identify individual skills and experience using text analysis. Thus, while not necessary to manually extract individual pieces of data, sections containing pertinent information were tagged.
- *Requests for Proposals*: Government RFPs in general include a title, objective, and

description, along with metadata such as a solicitation code and technical point of contact. The majority of RFPs follows a consistent format and therefore the sections of raw text identifying metadata, objectives and topic descriptions can be found.

Parsing and Storing Structured Data

Structured data comes in several forms, most notably in this case, spreadsheets and databases. Dates of hire, proposal and project information, and timecard data are all accessible via standard spreadsheets. Communication data from email, phone, and instant messaging are automatically stored in a database. Parsing this data and inserting it into the database is accomplished in a fairly straightforward manner. We describe the data extracted from these structured resources.

- *Hire Dates*: The dates of hire and, for former employees, termination, were maintained in a spreadsheet, and only minor algorithmic steps were necessary to ensure dates were formatted correctly for the database.
- *Timecards*: Timecard information for each employee is collected by an automated system. Monthly reports are generated as structured spreadsheets. We were able to extract specific pieces of data from the spreadsheets and insert them into the database. This data included each employee's name, the total number of hours logged for each month, and the total hours charged to each project during each month.
- *Proposal Information*: High-level information on submitted job proposals is stored in a structured spreadsheet. Included in this information are reference numbers, proposal titles, proposal managers, customer information, proposed budget, and likelihood of winning.
- *Project Information*: As with proposal information, high-level information on

each project at the company is stored in a structured spreadsheet. This information includes internal job numbers, the job number of the proposal which led to the project, the project title, the customer, the budget, begin and end dates, and the project manager.

- *Communications Data*: Data regarding e-mail, phone, and IM conversations is automatically logged and stored by the company's Information Systems division. This data includes the sender, recipient, timestamp, and size/duration of the communication. It does not, however, include information on the content of the messages or conversations. The IS system stores all of the logs in a database, so it was a simple matter of importing that information into the central TeamBuilder database.

Correlating the Database

After each data resource was formatted and inserted into the database, the next step was to correlate the information across different sources. Because the various data resources are maintained in different ways, by different people, and for different purposes, there is often no direct link between corresponding entities within different sources. Records between databases may not match for several reasons including: letter transpositions, omission of letters, misspellings, changes in personal information, use of nicknames, or even fraud (Cook & Cook 2003). Additionally, the manner in which entries are stored in different databases can make it difficult to match corresponding entities. For example, an employee's name might be stored as "John Smith" in one resource and as "JSMITH" in another.

In general, the task of matching multiple references to the same entity is referred to as record linkage. First introduced in the late 1950's, the theory of record linkage was later formalized by Felligi and Sunter (1969). In their article, Felligi and Sunter classify the decisions made when

comparing entity data as either a link (the data refers to the same entity), a non-link (the data does not refer to the same entity), or a possible link (manual review is required). An optimal linkage rule can then be defined which attempts to minimize the error when assigning links and non-links, as well as minimize the number of possible links for manual review. The actual rules and processes which are used to match the data can vary. The most straightforward approach to record linkage is a rules-based method. Here, common rules for matching entities are defined and then revised as exceptions occur. While this approach is relatively easy to implement, the number of rules and exceptions can grow very quickly. Standardization of the data can help mitigate many of the exceptions and rules (Welker 1993). Statistical and machine learning approaches to record linkage, such as Bayesian networks, offer more robust solutions, but require labeled training data (Gu 2003, Welker 2002).

For the TeamBuilder project, the decision was made to use a **mixed-initiative interaction** approach based on rules-based methodologies to correlating the data.

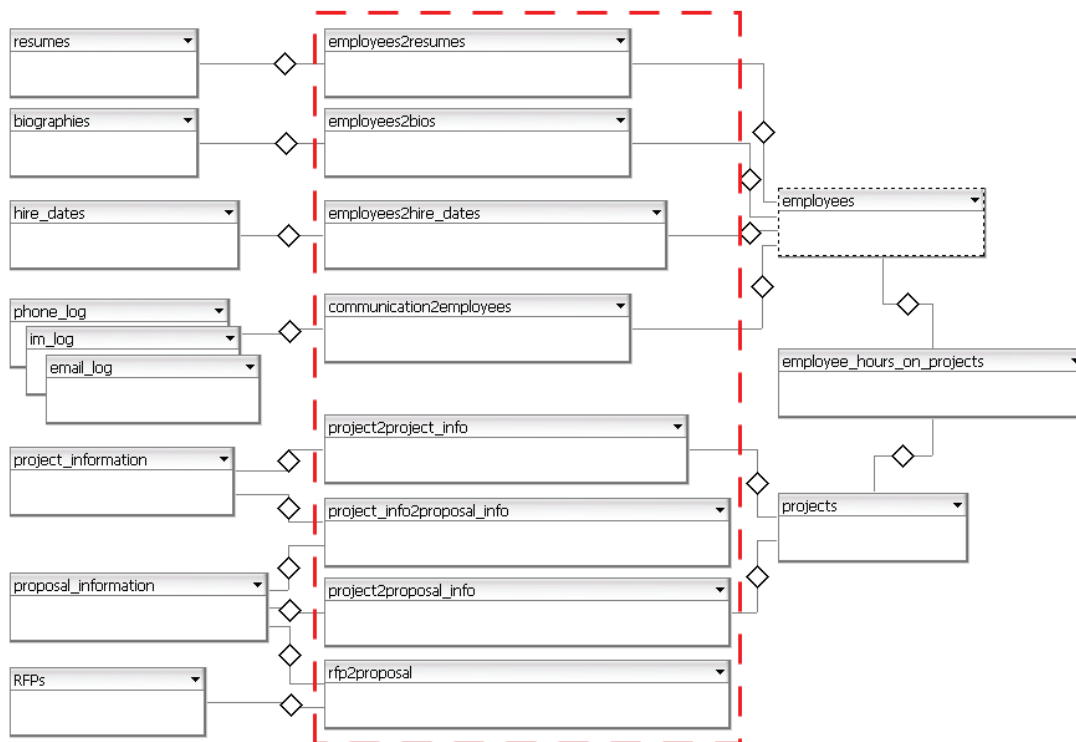
A mixed-initiative approach combines the strengths of the computer agent with the strengths of the human agent, applying each when appropriate (Hearst 1999). The goal of Day, et. al. (1997) in their mixed-initiative approach to language processing was to "transform the process of manual tagging to one dominated by manual review." Likewise with TeamBuilder, our aim is to push the burden for correlating entities onto the computer and leave the process of reviewing to the human user. This process takes advantage of the ability of the computer to quickly match similar entities across data resources along with any added knowledge that the user has which isn't explicitly formalized in either the computer program or data. For example, if an employee was to get married and changed her name, the computer may have difficulty linking the different name references to the same person. The human

operator can inject that knowledge in the review process to make the correlation.

This mixed-initiative approach was deemed appropriate for this application for a number of reasons. First, the time constraints placed on this portion of the project did not allow implementation of complex statistical methods for linking data. Instead, a rules-based approach was used in which individual data fields were matched across resources. The process of matching entities was accomplished through the text matching capabilities of the database. When possible matches were identified, the user was tasked with approving or rejecting the matches. In the case of incorrect matches, the user was asked to choose the correct match from the list of all entities. This approach is sub-optimal by Fellegi and Sunter’s definition, due to the fact that to the number of possible links identified is not minimized. But, once the

constraint of time was considered, this approach proved to be more desirable. Additionally, due to the size of the data (approximately 100 entities to be correlated), this particular mixed-initiative rules-based approach was manageable. With larger data sets this approach can still be applied, but caution is necessary to ensure that the burden on the user remains minimal. Using the computer to find possible matches and presenting the user with the ability to accept or reject the match is less time consuming than manually matching entities especially when dealing with large sets of data. However, when dealing with larger quantities of entities, optimizing the rules of automated matching to better identify positive links becomes increasingly important. Statistical methods are better suited to large data sets, and the time spent applying them is easily made up by the decrease in time spent during manual review.

Figure 2. Correlation tables in the TeamBuilder database



A Case Study in Data Mining

The first step to this approach to correlating tables within the database is to identify specific corresponding fields within each table. As stated previously, the employees and projects tables (derived from the timecard data) are taken to be the canonical repositories of employee names and project names and codes. Thus, all tables containing information about employees are correlated back to the employees table, and likewise for projects. A program was developed for the TeamBuilder project which allows the user to specify the rules for matching entities. For example, in the table for hire dates, employee names are stored in a single field by last name, first name. In the employees table, names are stored in separate fields for first and last. After the user has configured these values, the program finds the closest matches between the tables. Using a mixed-initiative approach, the results are displayed to the user who can then select the correct matches. Figure 2 illustrates the tables used for correlating different data sources within the database.

Privacy and Security Perspectives in Data Storage

In general, there are many techniques for protecting stored data. Based on the proposed application of the data and level of security required, one or more can be implemented to mitigate misuse of the data. Limiting access of the data resources to trusted users is a common approach (Clifton and Marks, 1996). At one end of the spectrum, access control mechanisms can be used to completely allow or deny access to the data. At a finer resolution, users can be given rights of use to only certain portions of the database to prevent full exposure of the information. Aggregation is another technique in which data is combined either within or across individuals. This can often be done in such a way as to prevent undesired uses of the data, while still enabling the planned applications (Verykios, et. al., 2004). Elimination

of unnecessary groupings can prevent unwanted information from being exposed. For example, when storing employee phone numbers, the use of the area code can be used to determine general locations of individuals. If the area code is not necessary for the intended application of the data, it should be eliminated from the database. Adding misleading entries to the database can obfuscate the data to unwanted queries. This must be done in such a way, though, so the intended access algorithms will only retrieve correct information. For example, suppose extra, fictitious people were added to a company phone book. It would still be possible to find the phone number of a known individual, but queries to find all individuals in a particular group would return additional fictitious people. This technique makes it more difficult to distinguish actual data points from the added noise. Perturbation of the data, through alteration by a known value or addition of noise, can limit access to users who know the perturbation scheme (Clifton and Marks, 1996).

To mitigate the risk of exposure of any information in TeamBuilder, it was deemed sufficient to implement access control procedures. A small subgroup of the development team was given admission to the database. Access was further limited by only giving users access to the portion of the data required for their research. While for some researchers it was necessary to allow full access to the data, others were, for example, allowed to view resumes, but denied privileges to the communication logs.

In addition to accidental exposure of private data, inaccurate correlation of disparate data resources can have undesirable social consequences. In the case of TeamBuilder specifically, if incorrect data is correlated to individuals it can bring about sub-optimal team assignments. Teams which are ill-equipped to face their missions, in terms of either teamwork or taskwork competencies, can lead to stressful social environments. A mixed-initiative approach to handling the data diminishes this risk by taking advantage of the

speed and capabilities of the computer along with human oversight.

Using the Data

Once the data has been organized into a common database, the TeamBuilder algorithms can operate on the data to provide measures of individual and team skills and competencies, and ultimately assemble near-optimal teams. The various algorithms which operate on the data include natural language processing, social-network analysis, and multivariate optimization. Each of these will be described along with other issues such as dealing with missing data. At this stage of the process, again, individual privacy is an important issue. Additionally, there are social implications involved in using TeamBuilder. While the system does incorporate a wide range of data, there are still exogenous factors that can influence the assignment of teams. We will discuss the steps taken to mitigate social concerns.

TeamBuilder analyzes three types of data. The first type of data is textual descriptions of the mission and of the skills of the employees: the taskwork skills needed and available. The second type of data is relationships between employees, as well as an employee's people skills as determined by social network analysis: the teamwork skills. Social network analysis (SNA) examines the social structure formed between individuals and applies mathematical techniques to derive measures such as *centrality* of leadership and *cohesion* of individuals or groups. The third type of data is the constraints with putting together any particular team due to availability, cost, and the like.

TeamBuilder is designed to be as unobtrusive, yet helpful, as possible. This requires that the analysis of these three types of data be done automatically with minimal input by supervisors and employees. In addition, the outputs of the system must make sense and be trusted by both parties. To enable this process, we take advan-

tage of a number of recent advancements in text analytics, social network analysis and multivariate utility assessment.

Assessing the Taskwork Skills of the Team

To match the skills of employees to the requirements of a mission, we use an unsupervised machine-learning technique called **Probabilistic Latent Semantic Analysis** (PLSA; Hoffman, 2001). PLSA produces a probabilistic multinomial model of the domain of interest – in this case the team and task-relevant characteristics of personnel. Although PLSA was initially developed as a word-focused tool for information retrieval that could function without dependence upon a dictionary, thesaurus, or a predefined taxonomy of concepts (Hofmann, 2001), it is a far more flexible methodology. Any type of discrete (i.e. count) data, such as elements in personnel bio-data, documents dealing with performance on projects, and more, can also be used in PLSA to assist in mapping the space of interest – here, taskwork skills. These skills are defined by “topics” which are simply distributions of words or other features.

As an example, consider the RFPs from a recent batch of Department of Defense Small Business Innovation Research program. These RFPs, which are generally one page, single-paced, represent the mission data. They describe the general objective to be achieved, a small background section and description of the problem, then the specific goals of the contract. For employee data, we have biographies (typically 1-2 paragraphs) and resumes of current and previous company employees and consultants. Both the RFPs and the bio-data were analyzed with the PLSA technique using a 40-topic model.

Table 2 illustrates the top ten most probable features from four of the topics in the model, in this case the features are words and their stems (e.g., materials, material, and materialization, would all

A Case Study in Data Mining

Table 2. Topics extracted from a corpora of text using PLSA

Topic 10	Topic 12	Topic 16	Topic 37
Materi	Health	Cultur	Human
Composit	Medic	Cultural	Factors
Materials	Diseas	Train	Factor
Composite	Care	Game	Design
Parachut	Clinic	Training	Interaction
Polym	Medical	Behaviors	Interfac
Textil	Clinical	Interactions	Interact
Composites	Blood	Behavior	Usabl
Fiber	Food	Cultures	Usability
Parachute	Ahlta	Interact	interface

be stemmed to “materi”). Even with this small amount of data some very clear topics emerge. Topic 10 is about composite materials for making parachutes. Topic 16 is about game-based training of cultural behaviors and interactions. Topic 37 is concerned with human factors and interface usability. Thus, PLSA is able to automatically extract the “gist” of a document, in this case an RFP or employee biography/resume. That is, after reading an RFP and being asked what it is about, the first few content words one would say would be in the topic, e.g., “Oh, it’s something to do with human factors and designing an interface.”

The topics create a layer of abstraction between the documents and the words in the documents. Thus, when comparing documents, matches can be made even when the individual words within the documents are mutually exclusive. For example, if the RFP uses only the terms “human factors design” and a person’s resume only uses “interface usability,” even though they refer to the

same skill, the two documents will be recognized as highly related because they both score high on Topic 37 in Table 2. Using these definitions of the topics, every RFP and biography/resume can be characterized by a “topic profile,” i.e., the extent to which the document is concerned with each topic. This information can then be used to determine the relevance of an employee to a particular RFP. This is achieved by directly comparing the topics of the employee with the topics of the RFP.

No one employee is likely to have all the necessary skills, so multiple employees will typically be required to achieve the goals of a mission. At the same time, one does not want skills duplicated between employees, therefore TeamBuilder aims to minimize the overlap between employees, while maximizing the match to the mission tasks.

Assessing the Teamwork Skills of the Team

In an organization, many employees are likely to have some of the same requisite skills, and teams of them can be created that have little overlap among those skill sets; but another aspect of team performance is how well the team members will work together to accomplish the task. Of the teams of people that do have all the skills required, which team would work together best and be most likely to succeed? To assess this, TeamBuilder takes advantage of a variety of analyses of the data.

Simple measures of success rates for individuals can fairly easily be derived directly from the data. Examining the percentage of projects an individual has been involved in that have had follow-on work can provide some insight into the past performance and success of that person. Combining measures such as this with other data, for instance percentage of total project hours worked by each team member, can further differentiate individual success from overall team success. More interesting information can be obtained by looking at other, “softer” aspects of an individual’s work behavior. For example, an employee’s “dependability” could be ascertained by looking at the frequency or speed with which they return emails to other people on the team when they are working on a project together versus when not (to account for more general friendship between them).

To further illustrate how measures of individual behaviors can be ascertained from the given data, consider the skill of **conflict management**. We define this as managerial experience in resolving potential disputes among team members. In the absence of data documenting past disputes, we operate under the hypothesis that functionally diverse teams (i.e. teams composed of individuals with different technical skills and backgrounds) are more likely to have conflict than teams composed of similar individuals. Conflict, in this

sense, is related to disagreements over technical approach, goals, etc., rather than personality conflicts. To measure the functional diversity of a team, some value of similarity between individual members is needed. PLSA provides a means of comparing team members. The resumes of individuals are used to train a PLSA model, and thereby comparisons can be made between each pair of team members. The functional diversity of the team is derived from these calculations. Teams with highly dissimilar members will be more functionally diverse.

To determine a manager’s experience in conflict management, we can therefore examine the functional diversity of all the teams he or she has managed in the past. To further refine this measure, the functional diversity of teams can be weighted by the project’s success. Thus, highly diverse teams that succeeded in their mission indicate higher conflict management skills of the manager than teams which were diverse but failed their mission.

Other kinds of specific teamwork skills may be required for a leader of the team. For example, a mission may require a leader that is good at direction-setting. That is, he or she is capable of determining the end state or specific end product. Information about this skill could be ascertained by a number of means such as looking at the number of previous projects they have led (a more general assessment), or looking at the number and length of email chains they initiate to project team members.

TeamBuilder will also assess measures of the team as a whole, such as cohesiveness using **social network analysis**. These analyses would, for example, examine the timecard data showing who has worked on what projects at the same time. High correlations between people indicate that they work on a lot of the same projects, and over time would suggest that they work well together. Measures of this type could also be weighted by the success of the projects to eliminate people

A Case Study in Data Mining

who work together a lot by circumstance but do not accomplish their tasks as well.

However, a fully-connected network of team members with previous experiences together may not be the ideal team design. Rather, a few individuals with many prior connections, and most with little or no prior connections, have been shown to enhance team functioning (Lazer & Katz, 2000). Further, a recent study on free-riding/social loafing in teams investigated the association between network ties and the amount of effort individuals invested in their teams. Their surprising finding was that although previous relationships with other team members was *not* related to team effort, the extent of common third-party relationships of team members (knowing the same people in common *outside* the team) was strongly related to team effort. The more people team members knew in common, the more effort they invested in their team work (Lazer & Katz, 2005).

Centrality, or the relative importance of individual nodes in a network, is an essential measure used in social network analysis. However, there is no single agreed upon conceptual definition or procedure to measure centrality (Freeman, 1979). We propose a measure based on email communications between team members. For each past project, we derive a value of centrality of each team member. By examining the total number of recipients on emails that an individual receives, the exclusiveness in communications for that individual can be determined. That is, we measure the degree to which team members send emails to an individual where he or she is the only (or one of a few) recipient. The centrality of a single team member can then be derived from his or her measures of communications exclusiveness with other members. Our hypothesis is that individuals who receive many communications addressed only to them play an important and central role on the team. This measure can be used to determine the functional leaders or the highly knowledgeable members of the team.

Taskwork skill assessment can be explained in a more straightforward manner, and gaps or overstatements of skills probably due to a lack of information by the TeamBuilder system that can be overcome. Teamwork skills and social connections however, may be more difficult to explain and thus less trusted by supervisors and employees. For example, no one will want to be labeled as less “dependable” than someone else, even if these measures reflect people’s (unstated) perceptions of others. To mitigate problems with such measures, supervisors will actually see this information only at the team level, e.g., the average dependability of one team versus another. In this manner, individuals maintain a certain (though not complete) amount of anonymity.

In addition, what relative weight should be put on dependability, or direction-setting or any other teamwork skill, versus taskwork skills is a matter of empirical research which we are conducting. This research will walk-forward through historical data to determine which of these parameters are best at determining a team’s success. In any case, as with the task skills, the supervisor will have the final say on who will be assigned to the team.

Assessing Constraints on the Team

After determining who, among those groups of people who have the right skills, would work well together and have a high probability of success, we must determine if these individuals are actually available to work at the right time, for the right amount of money, and at the least cost to other projects. These logistical constraints can be satisfied, and combined with the other two types of measures using the “**multivariate utility assessment algorithm**” (Levchuk, 2003) to assess the projected overall quality of the mission, based on projected quality of each task, past history of team members’ performance, and the relative cost in time, dollars and other projects’ probabilities for success.

The complexity of combining these different kinds of information is one reason for our development of TeamBuilder. No supervisor would be able to weigh all of these factors, let alone have access to the information on which they are based. At the same time, the reasons behind the ranking of teams given a mission must make some intuitive sense in order to be useful for supervisors to use. That is, if a supervisor had access to all the information, could figure out what weight each factor should have, then determine the costs and other constraints, then he or she should come to the same conclusion. As TeamBuilder is developed and deployed we will constantly monitor its “sanity” such that both supervisors and subordinates will trust its results enough to not be threatened.

There are several social implications to keep in mind resulting from this process. Depending on the accuracy of the data, TeamBuilder may not suggest perfectly capable individuals for a particular mission. To mitigate the consequences of the automatic assignments, we always present to the supervisor the list of potential team members and their skills/topics. The supervisor can at this point always add team members that might be appropriate, or remove those that are not. Thus, the system itself is merely a suggestion tool, performing triage, with ultimate responsibility up to supervisor. Allowing human override of the systems suggestions can have either positive or negative consequences, depending on the motives of the user. Ultimately, we feel that employees will be more comfortable with the final decision if it is derived from both the human and computer, rather than from one independently.

FUTURE TRENDS

Transitioning to Novel Organizations

Because the data mining conducted in this project was completely internal to the company, the legal

hurdles faced in collecting employee data were relatively few. Additionally, the types of personal data collected, such as resumes and timecards, were less invasive than other sources, such as performance reviews or pay scales. Transitioning a project such as TeamBuilder to new organizations presents a new set of problems.

Most organizations have strict data privacy policies in place and will allow access to different types of information. The way in which data is stored and used can depend on a number of factors, such as the size of the business or the type of business (government, military, private).

To cope with the different types and sources of data, TeamBuilder is designed to be adaptable to different situations. By creating a layer of abstraction between the teamwork and taskwork measures needed to drive the tool and the underlying data resources, the tool can handle changes in data availability. For example, measures of specific taskwork skills can be derived from resumes or biographies through natural language processing. If those sources are not available, other resources, such as surveys or skill assessment tests, can be substituted.

The greatest challenge to introducing TeamBuilder to novel environments will be the social, ethical, and legal issues involved with the particular organization. By bringing to light these concerns, strategies can be formulated to mitigate the risks involved in such a process. Ultimately, successful integration of a tool like TeamBuilder requires buy in from both the organization and the employees.

CONCLUSION

Traditional methods of assembling teams are based on manual matching of individual skills to the requirements of the mission tasks. These methods often do not consider the quality and mix of the teamwork and collaboration skills across the team members. This can lead to poor performance

A Case Study in Data Mining

even when each individual is technically qualified to accomplish his or her tasks. By automating the selection of teams and considering taskwork and teamwork competencies, the team members will be better prepared to tackle the mission goals in a collaborative manner. In this chapter we provided insight into the technical challenges of such an endeavor, as well as the social and ethical implications of automatically assigning teams.

The automated process of assigning teams requires large amounts of personal data about the potential team members. Using employee data can greatly enhance the quality of the product, but introduces many concerns. The ethical and social challenges associated with obtaining personal information are often more difficult than the legal concerns. Obtaining employee buy-in in order to use their data requires a careful approach and explanation of intentions. The technical issues inherent in dealing with disparate data resources are challenging, but can be dealt with through a systematic approach. This is required to ensure the integrity of the data and the soundness of automated assessments of employees and their assignment to projects. Through understanding of the risks involved in projects like the one described here, both individuals and organizations can benefit from strategic use of employee information.

ACKNOWLEDGMENT

This work was supported under Air Force contract FA8650-07-C-4510 with approval WPAFB 08-3088.

REFERENCES

Cannon-Bowers, J. A., Tannenbaum, S. I., Salas, E., & Volpe, C. E. (1995). Defining competencies and establishing team training requirements. In *Team effectiveness and decision making in or-*

ganizations, Guzzo & Salas (Eds.), Jossey-Bass, San Francisco, 333–380.

Clifton, C., & Marks, D. (1996). Security and privacy implications of data mining. In *Proc. 1996 SIG-MOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'96)*, Montreal, Canada., pp. 15-20,

Cook, J. S., & Cook, L. L. (2003). Social, ethical, and legal issues of data mining. In *Data Mining: Opportunities and Challenges*, Wang, J. (Ed.), Idea Group Publishing, Hershey, PA, 395-420.

Day, D., Aberdeen, J., Hirschman, L., Kozierok, R., Robinson, P., & Vilain, M. (1997). Mixed-Initiative Development of Language Processing Systems. *Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics*, 348-355.

Eddy, E. R., Stone, D. L., & Stone-Romero, E. F. (1999). The effects of information management policies on reactions to human resource systems: An integration of privacy and procedural justice perspectives. *Personnel Psychology*, 52, 335-358.

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Freeman, L. C. (1979). Centrality in Social Networks, Conceptual Clarification. *Social Networks*, 1, 215-239.

Fule, P., & Roddick, J. F. (2004). Detecting privacy and ethical sensitivity in data mining results. Appeared at *Twenty-Seventh Australasian Computer Science Conference (ACSC2004)*, Dunedin, New Zealand.

Gu, L., Baxter, R., Vickers, D., & Rainsford, C. (2003). Record Linkage: Current Practice and Future Directions. CMIS Technical Report No. 03/83, CSIRO Mathematical and Information Sciences, GPO Box 664, Canberra 2601, Australia.

- Hearst, M. (1999). Trends & Controversies: Mixed-initiative interaction. *IEEE Intelligent Systems*, 14(5), 14-23.
- Hoffman, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning Journal*, 42(1), 177-196.
- Kerr, N. L. (1983). Motivation losses in small groups: A social dilemma analysis. *Personality and Social Psychology*, 45, 819-828.
- Klimoski, R., & Jones, R. G. (1995). Staffing for effective group decision making: Key issues in matching people and task. In *Team effectiveness and decision making in organizations*, Guzzo & Salas (Eds.), Jossey-Bass, San Francisco, 292-332.
- Klimoski, R., & Zukin, L. (1999). Selection and staffing for team effectiveness. In *Supporting work team effectiveness: Best management practices for fostering high performance*, E. Sundstrom & Associates (Eds.), Jossey-Bass, San Francisco.
- Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, 7, 77-124.
- Lazer, D., & Katz, N. (2000). Putting the Network into Teamwork. Presented at the Academy of Management annual meeting, Toronto, Canada.
- Levchuk, G. M., Feili, Y., Pattipati, K. R., & Levchuk, Y. (2003). From hierarchies to heterarchies: Application of network optimization to design of organizational structures. *Proceedings of the 8th International Command and Control Research and Technology Symposium*, Washington, DC.
- Long, L. K., & Troutt, M. D. (2003). Data mining for human resource information systems. In *Data Mining: Opportunities and Challenges*, Wang, J. (Ed.), Idea Group Publishing, Hershey, PA, 366-381.
- Morgan, B. B., & Lassiter, D. L. (1992). Team composition and staffing. In *Teams: Their training and performance*, R. Sweezy & E. Salas (Eds.), Kluwer, Norwood, Mass., 75-100.
- Rouse, W. B., & Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin*, 100, 350-363.
- Rouse, W., Connon-Bowers, J., & Salas, E. (1992). The role of mental models in team performance in complex systems. *IEEE Trans. On Sys., man, and Cyber*, 22(6), 1296-1308.
- Saban, K. (2001). The data mining process: At a critical crossroads in development. *Journal of Database Marketing*, 8, 157-167.
- Salas, E., Burke, C. S., & Cannon-Bowers, J. A. (2002). What we know about designing and delivering team training. In *Creating, implementing, and managing effective training and development: State-of-the-art lessons for practice*, K. Kraiger (Ed.), Jossey-Bass, San Francisco, 234-259.
- Verykios, V. S., Bertine, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in Privacy Preserving Data Mining. *ACM SIGMOD Record*, 33(1), 50-57.
- Winkler, W. E. (1993). *Matching and record linkage*. Washington, D.C.: Bureau of the Census.
- Winkler, W. E. (2002) *Methods for Record Linkage and Bayesian Networks*. Washington, D.C.: Statistical Research Division, Bureau of the Census.

KEY TERMS

Centrality: A commonly used measure in social network analysis which determines the relative importance of a node within a network. There is no single, agreed-upon method to measure centrality; different methods are used depending on the application. In the case of building and

A Case Study in Data Mining

assessing teams, centrality within the communication network can be used to determine the functional leaders or knowledgeable members of the team.

Conflict Management: A measure of a leader's ability to successfully manage functionally diverse teams. When creating teams with members of highly varied backgrounds and skills, conflict management is an important managerial skill for resolving disputes related to goals and technical approaches. This is an example of a teamwork skill that can be derived from employee data.

Mixed-Initiative Interaction: An approach which combines the strengths of the computer agent and strengths of the human agent, applying each when appropriate. In record linkage, or the correlation of entities across data sources, mixed-initiative approaches can take advantage of the pattern matching abilities of the computer, leaving the task of reviewing the matches to the user.

Multivariate Utility Assessment: A mathematical optimization technique to assess the projected overall quality of the mission, based on projected quality of each task, past history of team members' performance, and the relative cost in time, dollars and other projects' probabilities for success.

Probabilistic Latent Semantic Analysis (PLSA): A statistical natural language processing technique for analysis of co-occurrence data within large corpora of text. PLSA finds the underlying topics, or "gist", of a document and can be used for searching or comparing documents.

Social Network Analysis (SNA): An analysis technique which examines the social structure formed between individuals and applies mathematical techniques to derive measures such as centrality of leadership and cohesion of individuals or groups

Taskwork Skills: Capabilities necessary to effectively complete all of the performance requirements for a particular task or project.

Teamwork Skills: The mix of attitudes, personality, and values that would optimize teamwork effectiveness and group cohesion among a particular set of individuals working within particular contexts.