IMPROVING THE QUALITY OF RFID DATA BY UTILISING A BAYESIAN NETWORK CLEANING METHOD

Peter Darcy Bela Stantic Abdul Sattar
Institute of Integrated and Intelligent Systems
Griffith University, Queensland, Australia
Email: {P.Darcy, B.Stantic, A.Sattar}@griffith.edu.au

ABSTRACT
Radio Frequency Identification (RFID) is a technology used to identify automatically a cluster of objects within a specified parameter. This technology has promised a means to cut cost of time and money in manual labor and to allow greater efficiency in numerous workplaces. However, there are various problems such as missed readings which hinder wide scale adoption of RFID systems. To this end we propose a system that utilises a Bayesian Network applied at a Deferred stage to impute and restore missed readings. Experimental results have shown that the optimal random threshold is 15% and that the DefBayNet method improves missed data restoration process when compared with the state-of-the-art method.

KEY WORDS
RFID Applications, Bayesian Networks, Data Imputation.

1 Introduction

Radio Frequency Identification (RFID), has been a hot topic of research in recent years due to its promises of wireless automatic group identification. The advent of this technology would usher in a new era where tasks such as super market check-outs and stock transferal can be performed easily with greatly increased efficiency. One of the main contributors as to why RFID is not employed in wide-scale deployment is that of unreliable and missed readings.

Different cleaning methods have been proposed in the past in an attempt to compensate for the faulty nature of the stored data but never to completely eliminate anomalies present in the data sets. We have found that each method proposed in the past have drawbacks, which hinders the RFID data set from providing maximum precision when querying the data. We believe that, by introducing a system which utilises a Bayesian Network to correct stored RFID data at a deferred stage of the capturing process, the majority of the missed readings can be restored.

To this end, we present DefBayNet, a deferred method of cleaning RFID data utilising a Bayesian Network. DefBayNet has been designed specifically to impute consecutive RFID missed readings in an effort to restore the data set with optimal integrity. We have run two experiments designed to demonstrate the cleaning abilities of DefBayNet, a Threshold Experiment and a Significance Experiment. The first experiment has been designed to find the optimal threshold to find both the highest precision and recall. The second experiment conducted was to compare the cleaning ability DefBayNet to that of an Cost-Conscious Cleaning Bayesian Network approach. From the results gathered, we have concluded that the optimal threshold for DefBayNet is 15% and that our approach outperforms the state-of-the-art method.

2 Background

Radio Frequency Identification (RFID) is a technology that has great potential to save time and money by offering a means to identify automatically a large sum of objects within a certain proximity. It does, however, have severe reliability problems in the form of unreliable readings and missed readings. Missed information in data sets are not unique to RFID applications as many other applications also suffer from this problem. A general method for solving this problem is to use Data Imputation to fill in the missing values with approximate values. However, to achieve this, an optimal threshold for probability must be found so that the Recall vs. Precision problem can be minimised providing both a high recall and precision rates.

2.1 RFID

Radio Frequency Identification, also known as RFID, is a system developed with the purpose of automatically identifying a group of objects within close proximity using electromagnetic relays of data. RFID has been integrated into applications since the 1940s [15]; in World War II, it was used to acknowledge a friendly aircraft on radars. The basic structure comprises a unique identification tag, a reader that interrogates the tag to extract the identifier and the RFID Middleware which is used primarily to filter out unwanted data. As a result, the data that gets recorded into the database is comprised of the identifier of the tag known as the Electronic Product Code (EPC), the identifier of the reader and a timestamp of the time of the reading.

The implementation of an RFID network is difficult due to the flaws in the passive architecture. These flaws include the high volumes of data, the low level data, the spatial and temporal aspects, and the data being extremely error-prone [1], [2] and [17]. There are three problems associated with the error-prone nature of raw RFID readings. These are duplicate data readings, unreliable data readings and missed data readings.

Duplicate data readings refer to any instance in which an RFID tag is captured twice. Unreliable Data Readings, also known as Noise, False Positives or Ghost Reads [3], are readings which are recorded on the reader but do not reflect reality. Missed Readings, also known as false negative readings, refer to any instances in which a RFID tag is supposed to be captured but, due a certain reason, is not. It is estimated that the RFID tags taken for any given situation only record 60% - 70% of the intended readings [4] and [6].

2.2 Data Imputation

Data Imputation is a process used in statistical applications to fill in missing gaps of data. When a data set has several known missing values present, the data is "imputed" through the use of different processes designed to derive an approximate value based on other data present. The goal of imputation is not always to correct the data set but, rather, to allow a plausible answer to be filled into the data source. This will allow the known data to be used in future analytical studies with minimal error as opposed to not being used at all [14].

2.3 Bayesian Networks

Bayesian Networks are a form of networks which assign a certain amount of probability to a set of variables that will then result in arriving at a conclusion [10]. The means in which we model this Network is in the mathematical equation:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | (X_1, ..., X_{i-1})) \qquad (1)$$

In the Equation 1, we can see that, in order to find the probability of X_i, we need to find the product of all the probabilities of X_1 to X_{i-1}. After the equation has been derived, a table is created and populated assigning a percentage for case of true and false for each child of *X* and parent of *X*. The system then multiplies each of the percentages together for each parent of *X* and ascertains the most likely candidate from the parent that received the highest percentage value.

2.4 Recall vs. Precision

The Recall vs. Precision is a well known problem that states that, in any given probabilistic situation, there will always be a trade-off between recall and precision [8]. With regards to this trade-off, recall is the amount of data that is to be fetched and precision is the accuracy of the data.

2.5 Related Work

Due to its unique properties and challenging scenarios, RFID data cleaning has become a frequently studied topic among the research community. Some methods which have been explored as a means to correct inaccurate RFID data at a deferred stage include: a Deferred Rule Based approach [13] and Probabilistic Integrity Constraints [7]. In past studies, Bayesian Networks have been employed to transform low level RFID data into higher level events, and in the process, attempting to correct anomalies present [11], [12]. Of these methods, we have been found one to be important in our study as it uses a Bayesian Network to also infer missed RFID readings.

The Cost Conscious Cleaning approach consists of an algorithm that has been designed to accept several rules. It examines the cost and accuracy of each cleaning method and finds the cheapest, most accurate cleaning solution. The study [5] also discusses the introduction of a Dynamic Bayesian Network uses a belief statement to determine if a tag is present based upon the immediate previous reading. We believe that conclusive evidence of the presence of a tag cannot be ascertained with maximum accuracy and that by using a deferred method of data cleaning we may yield a higher restoration rate through use of a Bayesian Network.

3 Deferred Bayesian Network Cleaning

To combat missing values within an RFID Data set, we present a Deferred Bayesian Network Cleaning method. This clenaing method has been designed to take advantage of a Bayesian Network to decipher the correct permutation of location recordings to fill in the missing gaps found in the data set. Unlike the Cost-Conscious Cleaning Bayesian Network method, [5], we have designed our system to evaluate the belief of a tag present based upon map data of the facility as well as past and future observations which can only be accessed from cleaning at a Deferred stage of the RFID cycle.

3.1 DefBayNet Structure

When designing DefBayNet, we had two goals in mind: a system which could be used to impute highly accurate data for missed RFID readings, and a system that allows an unlikely imputation to still have the amount probability it deserves to be chosen as an answer. Thus, we came up with the structure seen in Figure 1 which houses six processes. The inputs into this system are the Raw Data sets and the User, and the output being the Imputed Data Set.

The *User* will first define the *Bayesian Network* and *Threshold Percentage* to be used later on in the system. The *Analytical Process* is a simple program set up to extrapolate various characteristics from the Raw Data which are defined as important from the Bayesian Network.

At this point, the Threshold, Bayesian Network and Analytical Process all pass their findings to the *Imputation*

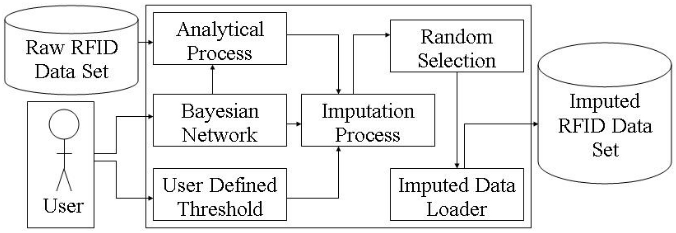

Figure 1. A high level diagram illustrating the processes and data flow present in the DefBayNet Architecture.

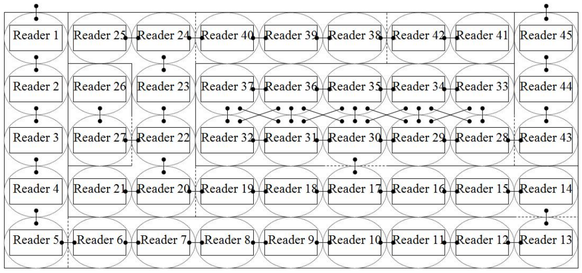

Figure 2. The floor map of a mock retirement village used within the experimentation of DefBayNet. The squares represent the readers, the circles represent the reader's reading range and lines with dots on either end symbolise access from one room into another.

Process which finds the different permutations. The Imputation Process then sends the found permutations to the *Random Selection* process which the system will make a weighted random choice as to which Permutation is correct based on the Bayesian Network output. A specified threshold is used to choose which values to randomly choose from. For example if there are the values of 10%, 6%, 4%, 2% and 1% and the system has a specified threshold of 25%, it will determine that of the percentages 10% and 6% are valid and will pick one of these values at a weighted random (i.e. taking a random number between 0 and 16, if the random value is between 0 and 6 the system will pick the Permutation that gained 6%, else it will choose the other Permutation). If no Permutation percentages lie within the threshold range, the system will pick the Permutation that achieved the highest percentage. The chosen Permutation is then passed to the *Imputed Data Loader* which loads this permutated data into the Data Set.

3.2 Experimental Scenario

The experimental scenario we have aimed for is utilising this process within a retirement village environment in which residents and staff are present and are all wearing RFID tags. The inspiration for this scenario is partly based on an article which discusses a hospital in Taiwan using RFID enabled wristbands to identify patients [16]. To this end, we have created a mock floor design as seen in Figure 2, and have created map data based on these floor designs to allow for higher accuracy when imputing data.

3.3 Database Structure

The database structure for the experimentations are based heavily on the Data Model for RFID Applications (DMRA) structure [9]. With regards to the experimentation we are populating and analysing the following entities: Reader, Object, Location, ReaderLocation and Observation. We have also created two extra tables which are used in the cleaning process called the Map Data and Bayesian Network tables. The Map Data table will be used as a look-up table to find how the rooms inside the retirement village are connected as shown in Figure 2. The Bayesian Network table will be used as a look-up table to find percentages of the permutations according to the row, column and truth values as recorded in Table 1. Additional tables have also been used for testing purposes to record and analyse DefBayNet's performance.

3.4 DefBayNet Network Table

The Bayesian Network Table for DefBayNet as displayed in Table 1 is the look-up tool used to find the user-defined percentage given to any situation. This table relies on seven truth or false variables and to determine the product of the percentages for each Permutation. Certain values have to be extracted in order to obtain the data necessary for the Bayesian Network, these values can all be seen in Figure 3 and include 'a' being the reader identifier two time stamps less than the first missed value, 'b' being the reader identifier one time stamp less than the first missed value, 'c' being the reader identifier one time stamp after the last missed value, 'd' being the the reader identifier two time stamps after the last missed value, 'n' being the number of consecutive missed reads and 's' being the number of readers that form the shortest path. We define the time stamp as the period that the readers scan for RFID tags.

The truth or false variables used in the Bayesian Network include whether or not: 'a' is equal to 'b' (a==b), 'b' is in the vicinity of 'c' according to the map data (b↔c), 'b' is equal to 'c' (b==c), 'c' is equal to 'd' (c==d), 'n' is equal to 's' minus two (n==(s-2)), 'n' is greater than 's' minus two (n>(s-2)) and 'n' is less than 's' minus two (n<(s-2)). With regards to the use of 's', we minus 2 from it as the shortest path will include the boundary values and, thus, these are not important.

Five Permutations of possible restored values are then generated from the data collected and each may be viewed in Figure 3. The first permutation will substitute each missed value with the 'b' reader value. The second permutation will substitute each missed reading with the 'c' reader value. The third permutation will find the shortest path, place it in the middle of the missed readings and substitute the reader closest for any missing values still present.

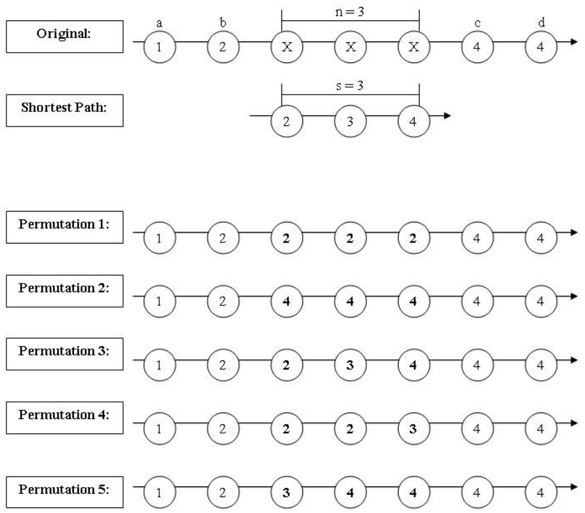

Figure 3. An example of a missed value situation in which DefBayNet would permutate five different routes.

The fourth permutation will take the shortest path and substitute it into the missed readings closest to 'b' and use the value 'c' as the replacement for any missed values still present. Permutation five will perform the opposite of permutation four.

To develop the table needed to produce the Bayesian Network, we needed to first come up with the general Equation to deal with missed values. Thus, Equation 2 reflects the mechanics we used to develop DefBayNet's Bayesian Network. In this Equation, P_i stands for Correct Permuatation, which is either Permutation 1, Permutation 2, Permutation 3, Permutation 4 or Permutation 5, and a, b, c, d, n and s are all variables found from the situation used in calculating the correct permutation.

$$\begin{aligned} P(P_i|a == b, b \leftrightarrow c, b == c, c == d, n == (s-2),\\ n > (s-2), n < (s-2)) = P(a == b) * P(b \leftrightarrow c)\\ *P(b == c) * P(c == d) * P(n == (s-2))\\ *P(n > (s-2)) * P(n < (s-2)) \end{aligned} \tag{2}$$

3.5 Assumptions

We have made three assumptions based upon the scenario we have created to ensure that DefBayNet runs at its optimal level:

- There are only missed read anomalies within the RFID data set.
- All readers have a time period in which they scan each time.
- The amount of time it takes for a person to walk to the next reader will be greater than the period of time each scan is taken.

4 Experimental Evaluation

For testing the Deferred Bayesian Network approach for significance in the field of RFID data cleaning, we have chosen to conduct two experiments. The first experiment is to run DEfBayNet on four different data sets setting the probability threshold at a different level each time. This is designed to find the optimal threshold level which maximises both recall and precision while cleaning. In the second experiment, we will test DefBayNet with the optimal found threshold against the Cost-Conscious Bayesian Network. From this experiment we hope to discern that DefBayNet improves the integrity of the data set more successfully than this current Bayesian Network approach.

4.1 Experimental Environment

DefBayNet stores all RFID data in the Data Model for RFID Applications (DMRA), all tables in DRMA and other additional tables along with processing commands written in procedure language PL/SQL were run with Oracle 10g SQL Plus. The computer used for experimentation is a Microsoft Windows XP machine running Service Pack 3, Intel(R) Pentium(R) 4 CPU 2.79GHz with 2.00 GB of RAM.

4.2 Experimental Data Sets

The Data Set used within the Experimentation are simulated RFID recordings of an imaginary RFID-enabled retirement village. This data set tracks the daily activities of four people within an elderly couple's house from 7:30am to 12:30pm resulting in 751 observations. The data generated will be contained in the Observation table of the DRMA architecture. The attributes that will be housed in this table include an EPC, reader identifier and timestamp. With regards to the first experiment, this dataset will have 65% of its records randomly deleted based upon the idea that only 60%-70% of RFID readings are recorded. For the second experiment we will use the same base data set, however, we created four sets from this data with different percentages of randomly deleted variables, specifically 50%, 60%, 70% and 80%.

5 Results and Analysis

As mentioned throughout this paper, it has been our intention to run two experiments to obtain two facts, the optimal threshold cut off and the amount of significance DefBayNet holds against the current use of Bayesian Network in RFID. To this end, the first experiment, which we have labelled "Threshold Experiment", uses four different thresholds to determine which percentage produces the most accurate imputation. The second experiment we labelled "Significance Experiment", uses the optimal threshold found in the last experiment to compare DefBayNet against Cost-Conscious Cleaning Bayesian Network.

Table 1. A Table depicting the configuration of the Bayesian Network. Please note that the ↔ symbolises that the two nodes are next to each other according to the map data.

Permutations	a == b		b ↔ c		b == c		c == d		n == (s - 2)		n > (s - 2)		n < (s - 2)	
	T	F	T	F	T	F	T	F	T	F	T	F	T	F
Permutation 1	70%	30%	70%	30%	90%	10%	30%	70%	20%	80%	10%	90%	20%	80%
Permutation 2	50%	50%	70%	30%	90%	10%	70%	30%	20%	80%	10%	90%	20%	80%
Permutation 3	10%	90%	40%	60%	30%	70%	10%	90%	90%	10%	20%	80%	50%	50%
Permutation 4	90%	10%	40%	60%	30%	70%	10%	90%	10%	90%	70%	30%	50%	50%
Permutation 5	10%	90%	40%	60%	30%	70%	90%	10%	10%	90%	70%	30%	50%	50%

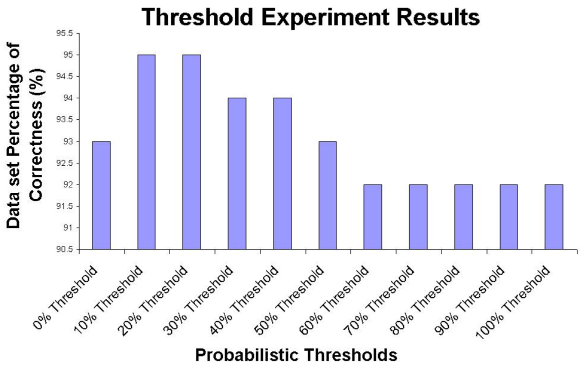

Figure 4. The results obtained from the Threshold Experiment plotted in a graph where the percentage of overall data set correctness vs. the different threshold settings.

5.1 Results

The first experiment we conducted was designed to test which threshold value was optimal for which to find a valid permutation. As such, we found the results when the threshold was set from 0% to 100% with every 10% in between tested. The results we gathered as displayed in Figure 4 show that the highest achieving thresholds are 10% and 20% obtaining a correct data set percentage of 95% and the worst thresholds being 60% to 100% who obtained 92% of correct data set percentage.

The second experiment we performed was designed to test the significance, if any, a Deferred Bayesian Network approach would make when tested against a Bayesian Network applied in Cost-Conscious Cleaning. To recreate an this Bayesian Network, we created a Bayesian Network that only gathered information obtained from the tags recorded directly before the current tag. The results we obtained from this "Significance Experiment" can be seen in Figure 5. The results have shown us that in each data set we tested on, DefBayNet would always beat the Cost-Conscious Cleaning method, also it appears that where there are more deleted values (i.e. the data set with 50% of its values deleted), DefBayNet obtains an even higher level of accuracy when examining the improvements from the Cost-Conscious approach.

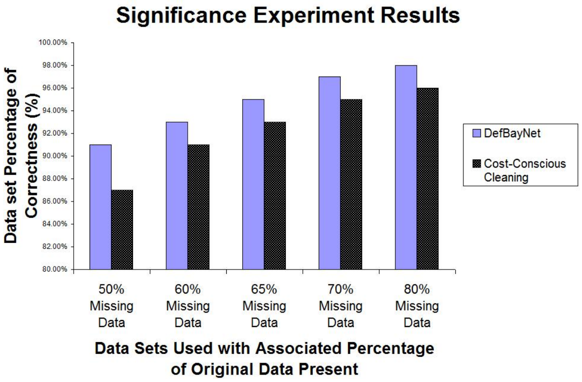

Figure 5. The results obtained from the Significance Experiment plotted in a graph where the percentage of overall data set correctness vs. the amount of random deletion each testing data set with the two series being the DefBayNet and Cost-Conscious Cleaning approaches.

5.2 Analysis

The advantages we have found of using a Deferred Bayesian Network approach when cleaning missed data are the higher reading rate and potential to provide an additionally higher accuracy when the data set has a high amount of consecutive missed data. The disadvantages of using our method are that the system has to be set up properly physically (the readers have to be placed precisely, the map data has to be recorded) and our system is required to be performed at a deferred stage, thus not providing a real-time solution to missed RFID reads. Finally, we would like to emphasies that the goal of Data Imputation, which we have based our model on, is not necessarily to provide 100% accurate data, but rather a highly accurate and plausible explanation as to what data could be present in missing gaps. We believe that DefBayNet performs highly accurate data imputation and at the very least, provides a plausible explanation as to what could be filled in the missing gaps of data.

6 Conclusion

In this paper, we introduced a novel cleaning method which utilises a Bayesian Network coupled with information which may only be obtained at a deferred stage of the RFID cycle to restore correctly the missed values inside an RFID data warehouse. In summary, we believe our study has made the following contributions to the field of RFID:

- We have shown that a Bayesian Network may be applied at a Deferred stage of the RFID capturing cycle and may be utilised to raw clean RFID data.
- We have found that 15% is the optimal cut off threshold when needing to find the right balance between recall vs. precision for a Bayesian Network applied at a deferred stage of the RFID reading cycle.
- In experimental evaluation, we have shown that DefBayNet performs more successfully than the current prominent application of Bayesian Networks where it is applied in the Cost-Conscious Cleaning appraoch.
- We have also proposed a method that will not only restore RFID data to the highest possible integrity, but also where it does not return a correct restoration, DefBayNet provides a plausible set of data.

With regards to future work generated from this research, we have intend to enhance DefBayNet's Permutations, Bayesian Network and modify it to be dynamic.

Acknowledgements

This research is partly sponsored by ARC (Australian Research Council) grant no DP0557303.

References

[1] Richard Cocci, Thanh Tran, Yanlei Diao, and Prashant J. Shenoy. Efficient Data Interpretation and Compression over RFID Streams. In *ICDE*, pages 1445–1447. IEEE, 2008.

[2] Roozbeh Derakhshan, Maria E. Orlowska, and Xue Li. RFID Data Management: Challenges and Opportunities. In *RFID 2007*, pages 175 – 182, 2007.

[3] Daniel W. Engels. On Ghost Reads in RFID Systems. Technical Report AUTOIDLABS−WP−SWNET−010, Auto-ID Labs, September 2005.

[4] Christian Floerkemeier and Matthias Lampe. Issues with RFID usage in ubiquitous computing applications. In Alois Ferscha and Friedemann Mattern, editors, *Pervasive Computing: Second International Conference, PERVASIVE 2004*, number 3001, pages 188–193, Linz/Vienna, Austria, apr 2004. Springer-Verlag.

[5] Hector Gonzalez, Jiawei Han, and Xuehua Shen. Cost-Conscious Cleaning of Massive RFID Data Sets. In *ICDE*, pages 1268–1272, 2007.

[6] Shawn R. Jeffery, Minos N. Garofalakis, and Michael J. Franklin. Adaptive Cleaning for RFID Data Streams. In *VLDB*, pages 163–174, 2006.

[7] Nodira Khoussainova, Magdalena Balazinska, and Dan Suciu. Towards correcting input data errors probabilistically using integrity constraints. In *MobiDE*, pages 43–50, 2006.

[8] Nodira Khoussainova, Magdalena Balazinska, and Dan Suciu. PEEX: Extracting Probabilistic Events from RFID Data. In *In ICDE*, 2008.

[9] S. Liu, F. Wang, and P. Liu. A Temporal RFID Data Model for Querying Physical Objects. Technical Report TR-88, TimeCenter, 2007.

[10] Daryle Niedermayer. An Introduction to Bayesian Networks and their Contemporary Applications [online]. niedermayer.ca, December 1998. Available from: <http://www.niedermayer.ca/papers/bayesian/index.html> [Accessed: 2nd October 2008].

[11] Donald J. Patterson, Lin Liao, Dieter Fox, and Henry A. Kautz. Inferring High-Level Behavior from Low-Level Sensors. In *Ubicomp*, pages 73–89, 2003.

[12] Matthair Philipose, Kenneth P. Fishkin, Mike Perkawitz, Donald J. Patterson, Dieter Fox, Henry Kautz, and Dirk Hahnel. Inferring Activities from Interactions with Objects. *Pervasive Computing*, pages 10–17, December 2004.

[13] Jun Rao, Sangeeta Doraiswamy, Hetal Thakkar, and Latha S. Colby. A Deferred Cleansing Method for RFID Data Analytics. In *VLDB*, pages 175–186, 2006.

[14] Joseph L. Schafer and John W. Graham. Missing Data: Our View of the State of the Art. *Psychological Methods*, 7(2):147–177, 2002.

[15] Harry Stockman. Communication by means of reflected power. In *IRE*, pages 1196–1204, 1948.

[16] Claire Swedberg. Hospital Uses RFID for Surgical Patients [online]. RFID Journal, July 2005. Available from: <http://www.rfidjournal.com/article/articleview/1714/1/1/> [Accessed: 5th June 2008].

[17] Fusheng Wang and Peiya Liu. Temporal Management of RFID Data. In *VLDB*, pages 1128–1139, 2005.