Identifying Relationships in Transactional Data

Melissa Rodrigues[1], João Gama[1,2], and Carlos Abreu Ferreira[2,3]

[1] FEP - University of Porto, Porto Portugal
[2] LIAAD-INESC TEC, Porto, Portugal
[3] ISEP - Polytechnic Institute of Porto, Porto, Portugal

Abstract. *Association rules* is the traditional way used to study market basket or transactional data. One drawback of this analysis is the huge number of rules generated. As a complement to *association rules*, *Association Rules Network (ARN)*, based on *Social Network Analysis (SNA)* has been proposed by several researchers. In this work we study a real market basket analysis problem, available in a Belgian supermarket, using *ARNs*. We learn *ARNs* by considering the relationships between items that appear more often in the consequent of the *association rules*. Moreover, we propose a more compact variant of *ARNs*: the *Maximal Itemsets Social Network*. In order to assess the quality of these structures, we compute *SNA* based metrics, like *weighted degree* and *utility of community*.

Keywords: Social Network Analysis, Association Rules Network.

1 Introduction

Every day the consumers satisfy their needs with the acquisition of products and services that they choose according to factors such as the price, the brand and the quality. To study the behavior and preferences of the consumers *market basket analysis* can be a powerful tool to help food chains, recommendation systems and other businesses to promote their products or services.

The traditional way to find relationships among products available in transactional databases is to run an *association rule miner* [1] and obtain *association rules* that represent valuable knowledge. The *association rule miner* algorithm generates a set of rules and each rule is associated with one, or more, interest measures, like support and confidence. The problem with this approach is the huge number of rules found by the *association rule miner*, most of them are redundant and uninteresting. A more recent framework used in market basket analysis is to use *Social Networks Analysis (SNA)* [2]. Using this framework a *social network of products* is built [13]. Social networks of products are graph structures where vertices can be products and edges represent products bought together in the same transaction. These networks can represent valuable relationships among products.

In this work we study the problem of identifying meaningful relationships available in the transactions database of a Belgian supermarket. We use two

different strategies to study this data. One strategy is to represent our data in a *global social network* (GSN), that includes all products, and use this representation to extract interesting information about the *best seller products*. We extract some interesting information about the products and network by computing statistical measures according to the vertex and according to the network. Moreover, we choose the *weighted degree* measure to find the *best seller products* [9]. The problem with this approach is that we have a large number of products and get highly complex networks that can not represent the motivation for two products appearing in the same transaction. To address this issue of our *GSN* we developed a new pipeline methodology that is grounded in the *Association Rule* analysis. First, we run an *association rule miner* to obtain the most interesting association rules. Then, and also to reduce the size of the ruleset, we put the focus on the most interesting products [13] and learn meaningful *Association Rules Networks (ARNs)* [12]. The most interesting products are the ones that appear more often in the consequent of the association rules found by the *association rule miner*. Moreover, we compute the same metrics that we use in the *SNA,* plus some communities measures used to detect and study *communities* [13] available in the obtained *ARNs.* We obtain some interesting results but we would like to study a more compact way to represent only the more meaningful relationships among products. Thus, we introduce in this work the *Maximal Itemsets Social Network (MISN)*. First we find the *maximal itemsets* [14] and then we generate the MISN.

The paper is organized as follows. In Section 2 we present some related work and concepts. In Section 3 we describe the methodology that we use to study our data and present the obtained results. Last, in Section 4, we conclude and present some work that we will develop in the near future.

2 Related Work

Identifying relationships in transactional data is the primary focus of *market basket analysis* [4, 12, 13].

According to [4], finding the hit-list of products in transactional data can be done by using an *association rule* framework but it is required to integrate the search for frequent itemsets within a microeconomic model. Otherwise we will obtain meaningless *association rules*. In this work the authors study a dataset of transactions acquired from a fully-automated convenience store and obtained interesting association rules. The authors also show that by exploring *frequent itemsets*, it is possible to identify the cross-sales potential of product items and use this information for better product selection.

Another work that discovers meaningful relationships among the products is [13]. In this work the authors model the data as a product network and present a new metric to study *communities*. To find interesting *communities* the authors introduce the *utility of community* metric. According to these authors, the set of *communities* and the *ARN* structure, along with the actual list of *association rules*, can provide important insights on the supermarket customer behavior. By

exploring data collected from a convenience store they found several interesting communities. For instance, the community associated with the highest value of the *utility of community* metric shows that chips and salsa are complementary products and that people often buy these two products together.

3 Methodology and Analysis of the Results

In this section we present the methodologies that we use to identify meaningful relationships available in the *Retail* database [4] while presenting and analyzing the results that we obtained. First, we represent the entire transactional data using a *social network*(*GSN*) and study some statistical measures according to the vertex and according to the network. Second, grounded in the *association rule* framework we present a sequence of steps that uncover interesting relationships. We start by finding the most interesting *association rules* using the *Apriori* algorithm [1]. Then, to reduce the ruleset and network complexity, we select the *association rules* having each one of the five products that appear more often in the consequent of the ruleset and use these findings to generate five *ARNs*, one *ARN* for each one of the five most frequent products. We analyze the obtained *ARNs* using the same statistical measures that we use to study the *social network* plus measures suitable to study *communities*. Next, we introduce a more compact representation based on the *maximal itemsets* [14] that we found in the entire *Retail* dataset, we introduce the *MISN*. To study this later *social network* we use the same metrics that we use to study the ARNs.

Concerning to the *Retail* dataset [4], this dataset registers 250 transactions and 1209 products sold in a small Belgium supermarket during 5 years and a half. We do not have assess to the identification of any of the products. Each product identity is mapped into an integer number and here we work only with the coded data.

3.1 Social Network Analysis

In this section we use SNA to represent the entire set of *Retail* transactions into a *social network* of products [13]. In the next sections this network is called *GSN*. In this network representation, each product defines a vertex. And every two products appearing in the same transaction are represented by an edge. Using this mapping strategy we obtain the *social network* represented in Figure 1. This network has 1209 vertices and 12257 edges. We use *Gephi software* [15] to draw the *social network* represented in this figure. To characterize, understand and explain the structure of this *social network* we compute and analyze a set of statistical metrics. We compute metrics for the vertices: *degree, weighted degree, betweenness and eigenvector centrality* and compute metrics for the network: *density, path length, number of the shortest paths, diameter and agglomeration coefficient* [2, 10]. Moreover, due to the poor viewing, we use the *Fruchterman Reingold* algorithm [7] to improve the networks' perception.

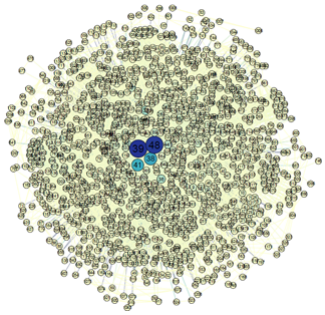

Fig. 1. Social network of *Retail* database

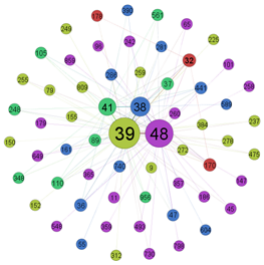

Fig. 2. *MISN* of *Retail* database

In Table 1 we present the average value of each metric that we use to analyze the vertices of the generated *social network,* the one presented in Figure 1. We can see that each vertex is connected to approximately 20 vertices. Regarding the *weighted degree*, the total weight average of the edges in each vertex is 21.838. On the other hand, the average value of the metric *betweenness* shows that a given vertex appears 810 times in the shortest paths of the *Social Network.* Last, by inspecting the average *eigenvector centrality* we can see that the vertices are connected about 0.057 among them.

In Table 2 we present the metrics that we use to analyze our *social network.* We see that we have a low *density* (0.017) network. This means that the network is far to become complete (complete networks have density equal to 1). The analysis of the metric *average length path* indicates that the average distance between all vertices pairs in the network is 2.364. Moreover, we can say that there exist 1436418 different shortest paths. Furthermore, we can see that the longest distance between two vertices, *diameter*, is 5 edges. The network presents a high *agglomeration coefficient* value (0.886) which suggests a tendency of the network to form cliques.

In Table 3 we present the five products with the highest *weighted degree* in the *social network.* The *weighted degree* is defined as the sum of the weights of all edges connected to a particular vertex [9]. Using this metric, we identify the five *best seller products*: **39**, **48**, **38**, **41** and **32**. Then, we represent the network of each product using *self-social networks.* These individual structures represent only the direct edges among each one of the best seller products and all the others products. We analyze these structures and get an *agglomeration coefficient* equal to zero. This shows that the *best seller* products play a central role and they can be seen as a communication channel among products [8]. Moreover, we get for all five *self-social networks* a *diameter* equal to 2.

Table 1. Statistical measures by vertex

Av. degree	Av. weighted degree	Av. betweenness	Av. eigenvector centrality
20.276	21.838	810.300	0.057

Table 2. Statistical measures by network

Density	Av. length path	Nr of shortest paths	Diameter	Av. agglomeration coefficient
0.017	2.364	1436418	5	0.886

Table 3. Products with highest *weighted degree*

Products	Weighted degree
39	1155
48	1092
38	618
41	556
32	233

3.2 Association Rules-based Approaches

In this section, grounded in the *association rule framework*, we use methodologies to reduce the complexity of the social networks. We learn an *ARN* for each one of the most interesting products. Moreover, we introduce the *MISN,* a more compact representation of the entire *Retail* database that uses the *maximal itemsets*.

Association Rules We use association rules to explore the relationships between items frequently purchased together. These rules have two parts, the *antecedent* part and the *consequent* part. For instance, in a supermarket dataset we can obtain the following rule: $\{cake, bread\} => coffee$. This rule represents a partial implication that means that people buying cake and bread also buy coffee. To find *association rules* there exist a wide number of algorithms that we can use [1, 6]. In this work we will use the *Apriori* algorithm [1] to find the most interesting rules. The *Apriori* algorithm can be parameterized to find both the *association rules* and the most *frequent itemsets* [1]. Moreover we can use a naive post-processing strategy to find the subset of *maximal itemsets*. An itemset is *maximal* if it is *frequent* but none of their proper *supersets* is *frequent* [14].

Typically, when we run an *association rule miner* to find either the *frequent itemsets* or the *association rules* on a database of transactions, we get a huge number of uninteresting and redundant patterns. If we increase the minimum *support* or *confidence* thresholds we get a small number of patterns but we can fail to detect interesting itemsets or rules having a support or/and confidence value lower than the user-defined values. To address this issue we can use low support and confidence thresholds and compute additional metrics like *conviction, cosine or lift* [6] to select the most interesting *association rules*. In this work we compute the *lift* metric of each *association rule* to find rules having negative or positive association. In Table 4 we present a subset of association rules that we obtained by running *Apriori* algorithm to find *association rules*. We use the implementation available in *arules* package of the *R Project software* [16]. We set the *support* threshold value equal to 0.01, the *confidence* threshold equal to 0.50 and the minimum *lift* value of each rule equal to 2. With this setting we obtained 115 rules. Without the *lift* constrain and using the other two parameters we get 288 rules, i.e., by using *lift* we prune 173 rules.

Association Rules Networks Here we learn an *ARN* for each one of the top 5 most frequent products available in the consequent of the 115 association rules presented in the previous section. Then we analyze each one of the *ARNs.*

Table 4. Three association rules that we found by setting $lift > 2$

Rule Nr.	Antecedent		Consequent	Support	Confidence	Lift
2	281	=>	38	0.012	1	3.676
274	36,39,48	=>	38	0.016	1	3.676
288	36,38,39,48	=>	41	0.012	0.75	3.074

According to [12], an *ARN* can show the direct and indirect associations among the products. Each *ARN* is a unique direct hypergraph that is associated with only one target-product. Each rule is represented by the directed edge where the itemsets in the rule antecedent are the source-vertices and the target item in the consequent of the rule is the destiny-vertex. These authors present the following sequence of steps to learn an *ARN:* **1)** Given a database D and the minimum support and confidence threshold, find all *association rules* using an algorithm such as *Apriori*; **2)** Choose a frequent item z that appears in the consequent of a ruleset and built the direct hypergraph that flows to this target-vertex z; **3)** Prune the constructed hypergraph of opposite hypercycles and hyperedges. The resulting hypergraph is an *ARN.*

We analyze the *ARNs* using the metrics described in Section 3.1, we compute metrics according to the vertex and according to the network. However, due to space limitations we can only say that we obtained lower *degree* and *weigthed degree* values when we compare with the values that we obtained in the *GSN* (see Section 3.1). They express few relationships but still get a value higher than the ones obtained when analyzing the *social self-networks.* We also find that the *density* is high in the *ARNs* as they represent just the most interesting rulesets. As *ARNs* have less relationships compared to the relationships of the *GSN,* their *agglomeration coefficient* is lower, which means a lower tendency to form cliques.

Here we also study the *communities* in each *ARN.* A *community* is a group of vertices densely connected that has high concentrations of edges connecting vertices within the group and has low concentrations of edges connecting to other groups. The metric used to detect this property is called *modularity* [11, 5]. *Modularity*, Q, values range between -1 and 1:

$$Q = \sum (e_{II} - a_I^2) \tag{1}$$

where, e_{II} is the fraction of edges that join the vertices to other vertices in the community I and a_I is the fraction that remains in the community I.

To evaluate these groups, a new measure called *utility of community* $U(G_i)$ was proposed by [13]. This measure includes two parts: *information* and *information density* and its range is between 0 and 1. *Information*, $I(G_i)$, is the sum of the weights of the *intra-community* edges, $I(G_i) = a_0 + \sum P(p_1|p_2)$. *Information density* $D(G_i)$ give us the information of the vertex i in the graph G_i, in a given *community*, in other words, its *weighted degree*, $D(G_i) = \frac{I(G_i)}{|V_i|}$. This way we can define the *utility of the community*:

$$U(G_i) = \frac{2I(G_i)D(G_i)}{I(G_i) + D(G_i)} \tag{2}$$

Table 5. Statistical measures by vertex of the *MISN*

Av. degree	Av. weighted degree	Av. betweenness	Av. eigenvector centrality
3.806	6.484	34.371	0.199

Table 6. Statistical measures by network of the *MISN*

Density	Av. length path	Nr. of short paths	Diameter	Av. agg. coefficient	Modularity
0.062	2.127	3782	3	0.534	0.255

There are several algorithms to detect *communities*, so in this work we use an algorithm of *modularity optimization*, the *Blondel Algorithm* [5]. This algorithm has two phases. Consider a weighted network where each vertex is a community, the first phase is to calculate the *modularity gain* of all neighbors. If a vertice has a *modularity gain* (ΔQ) higher in a neighbor community, it will be allocated to this community. This process continues until there is no improvement. After creating the network, the second phase is to apply again the first phase of the algorithm to the weighted network [3, 5].

In this work we follow this methodology to discover *communities* in the five *ARNs*. The *ARNs* correspond to the products **38**, **41**, **48**, **32** and **36** and we analyze each network by computing the same metrics that we used in Section 3.1. Moreover, we analyze communities available in each *ARN* by computing some special purpose measures: *modularity, representation percentage and utility of community*.

ARNs provided us to discover expressive relationships between some popular products. In these informative *ARNs* we could imagine a similar need, such as complementary, or products purchased in the nights, for instance. According to [13], *ARNs* and *association rules* failed to find relationships that fall outside the specified support and confidence thresholds.

Maximal Itemsets Social Network (MISN) Here we introduce the *MISN* to get a more compact network representation than the *GSN* that we found in Section 3.1. We believe that the *MISN* will help us to discover meaningful relationships in the entire transactions database. To obtain the *MISN* we run the *Apriori algorithm* to find all *frequent itemsets* and then use a naive strategy to find the *maximal itemsets*. Using this strategy we found 117 *maximal itemsets*. Then, we translate the *maximal itemsets* into an adjacency list. Next, we use the adjacency list to generate the *MISN*. Finally, we analyze the statistical measures of the *MISN* and detect the *communities*. We also compare the *MISN* communities, where the target products appear, with the corresponding *ARNs* communities.

In Table 5 we present *MISN* vertex statistical measures. The metric *degree* shows that the average number of edges connecting to a vertex is approximately 4 edges. This table also shows that the average *weighted degree* is 6.484. Concerning the measure *betweenness*, that reflects the number of times that a given vertex appears in the *shortest paths* of the network, we get an average value of 34. In the last column we present the metric *eigenvector centrality*. We can see that each vertex is connected to 0.199 of the vertices.

Table 7. Communities, products, representation percentage and utility in *MISN*

Communities	Products	Rep. percentage	Utility
1 (violet)	48,859,798,96,260,359,147,357,179,365,548,186,258,242,101,45,493,649,730,11,65	33.87	0.385
2 (light green)	39,79,152,384,150,259,155,272,225,312,9,249,475,278,237,809,255	27.42	0.269
3 (dark green)	248,41,348,956,37,89,105,561,110	14.52	0.235
4 (red)	178,32,170	4.84	0.167
5 (blue)	38,55,604,589,281,161,390,36,740,441,286,47	19.35	0.326

Relatively to the metric *density* presented in Table 6 we get a density of 0.062, which reflects that the *MISN* is far to be complete. The measure *average length path* reveals that 2.127 is the average distance between all the pairs of vertices. In the total, 3782 is the sum of all *shortest paths* among each pair of vertices. The longest distance between any two vertices is obtained using the *diameter* metric, which in this case is of 3 edges. The average *agglomeration coefficient* is 0.534 and this value shows the tendency to cliques formation, which are complete subgraphs where any two vertices are connected at least by a edge. In this case, we can see in Figure 2 a central clique, composed by the products **41, 38, 39** and **48**, for instance, which correspond exactly to the products with higher *weighted degree* in the *MISN* and in the *GSN*.

Here, we also study the five products that have the higher *weighted degree* (**39**, **48**, **38**, **41** and **32**). Suppose that these products can be *bread*, *apples*, *cheese*, *coffee* and *sugar*. We imagine these because they satisfy needs in all meals, like *bread* can be necessary for the breakfast, lunch and dinner and their function can be complementary with other products, like *cheese* with ham, a meal composed by rice, tomatoes, beef, *apples* and *sugar*.

Beyond the representation and statistical metrics, it is important to analyze the communities. In this sense, we apply again the *Blondel algorithm*. This way we can obtain the number of discovered *communities*, the *representation percentage* in the entire network and the *utility of community*. The *MISN modularity* has the value of 0.255, so it has a significant community structure [10] and we identified 5 communities, as we can see in Figure 2.

The *utility of community* helped to quantify the importance of each community in the *MISN,* according to the weights of the edges and the vertices' *weighted degree*. This is the contribution that the groups provide to the *social network*, excluding the illusory effect of the representation percentage, that just take into account the number of the vertices comprised in each *community*. This way, we sort the *communities* of the *MISN* using the *utility* metric: the *community* **1** with 0.385, the *community* **5** with 0.326, the *community* **2** with 0.269, the

Table 8. Community of the target-product, products, representation percentage and utility of the community of each *ARN*

Chosen ARN	Com. of the target-product (ARN)	Products	Rep. percentage	Utility of the community
38	1	**38**,55,281,32,589,604,170,740	40	0.242
41	2	**41**,38,441,105,390,32,37,248	50	0.504
48	1	**48**,798,357,11,859,96,147	35	0.24
32	1	**32**,178,39	100	0.889
36	1	**36**,740	50	0.40

community **3** with 0.235 and finally the *community* **4** with 0.167. In the Table 7 we can see that the second community (second row) in the *MISN* has the second higher *percentage of representation*, but its *utility of community* (0.269) is lower than the *utility of community* of community five (0.326). Therefore, communities that include a higher number of products do not always correspond to the communities with highest *utility of community*.

Comparing the communities of the *MISN* network with the communities of the five *ARNs* (products **38, 41, 48, 32** and **36**), we found (see Table 8) that: 50% of the products contained in the community of the *38*-ARN that contains the product **38** are also contained in community **5** of the *MISN* (see Table 7), that also contains the product **38**; 37.5% of the products contained in the community of the **41**-ARN that contains the product **41** are also contained in community **3** of the *MISN,* that also contains the product **41**; 66.7% of the products contained in the community of the **32**-ARN that contains the product **32** are also contained in the community **4** of the *MISN;* 100% of the products that appear in the ARN community of the **48**-ARN that contains the product **48** and in the community of the **36**-ARN that contains product **36**, appear, respectively, in the communities **1** and **5** of the *MISN*. Thus, we concluded that the *MISN* is a good approach to discover meaningful relationships and that most of the *communities* of the chosen products in the *ARNs* are well represented in the *MISN*.

4 Conclusions

In this work we explore different representations of our supermarket transactional data and discovered very interesting and meaningful relationships. We start by building a *social network* of products, our *GSN,* and found that this representation generates a highly complex network. This can be the result of representing all products bought together, including the ones that were bought without a common motivation. Then we use the *association rules* to explore the relationships between the products. The problem with this approach is the huge number of *association rules* found by any association rule miner. Then we introduce the *lift* measure to get a small and more interesting set of rules. This way, we get some interesting rules but we need to search for a global overview on the products. Thus, we select the top five frequent products in the consequent of the lift-pruned rules and generate five *ARNs*. We analyze each one of the five *ARNs* using a set of metrics and search for the *best seller products*, the products having the highest *weighted degree*. Moreover, we found that the *agglomeration coefficient* has lower values for the *social self-networks* when compared with the values that we obtained in the *ARNs*. This is explained by the *social self-networks* representation structure. *Social self-networks* represent the edges between the vertices and the *best seller vertex*, not taking into account possible edges between the vertices. Last, we use the *maximal itemsets* that we find in the entire dataset to generate a more compact network, the *MISN*. Overall we can say that the *best seller products*, according to the measure *weighted*

degree, computed from the *GSN*, do not correspond to the products that appear more often in the consequent of the *association rules*, but to the products with higher *weighted degree* in the *MISN*.

In the future, we will search for other interest measures for the *association rules* to obtain high interesting association rules and explore the *cliques* available in the *MISN* to group products that share a significant level of connections.

Acknowledgments: This work is funded by the ERDF - through the COMPETE Programme and by National Funds through the FCT Project KDUS.

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. Proc. of 20th Intl. Conf. on Very Large Data Bases, 487–499, Santiago, Chile (1994)
2. Albert, R., Barabasi, A.: Statistical mechanisms of complex networks. Reviews of Modern Physics, 74, 47–97 (2002)
3. Blondel, V.D., Guillaumem J., Lambiotte, R., Lefebrvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 10 (2008)
4. Brijs, T., Swinnen, G., Vanhoof, K., Wets, G.: Using Association Rules for Product Assortment Decisions: A Case Study. Proc. of the 5th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining, San Diego, CA, USA, 254–260 (1999)
5. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press (2010)
6. Faceli, K., Lorena, A.C., Gama, J., Carvalho, A.: Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina. Livros Técnicos e Científicos (2011)
7. Fruchterman, T., Reingold, E. M.: Graph Drawing by Force-Directed Placement. Software: Practice and Experience, 21(11) (1991)
8. Kretschmer, H., Kretschmer, T.: Application of a New Centrality Measure for Social Network Analysis to Bibliometric and Webometric Data. 1st IEEE Int. Conf. of Digital Information Management, 199–204 (2006)
9. Lopez-Fernandez, L., Robles, G., Gonzalez-Barahona, J. M.: Applying Social Network Analysis to the Information in CVS Repositories. 1st International Workshop on Mining Software Repositories (MSR), 101–105 (2004)
10. Newman, M. E. J.: The Structure and Function of Complex Networks. Society for Industrial and Applied Mathematics Review, 45(2), 167–256 (2003)
11. Newman, M. E. J.: Fast algorithm for detecting community structure in networks. Physical Review E, 69, 066133 (2004)
12. Pandey, G., Chawla, S., Poon, S., Arunasalam, B., Davis, J.: Association Rules Network: Definition and Applications. Statistical Analysis and Data Mining, 1(4), 260–279 (2009)
13. Raeder, T., Chawla, N.: Market Basket Analysis with Networks. Social Network Analysis and Mining, 2(1), 97–113 (2011)
14. Gouda, K., Zaki, M.: Efficiently mining maximal frequent itemsets. Proc. of the International Conference on Data Mining, 163–170 (2001)
15. Bastian M., Heymann S., Jacomy M.: Gephi: an open source software for exploring and manipulating networks. AAAI Conf. on Weblogs and Social Media (2009)
16. R Core Team: R: A Language and Environment for Statistical. R Foundation for Statistical Computing (2012)