Mining Consumer Complaints*

Extended Abstract†

Cary Sullivan
University of Colorado Boulder
Cary.sullivan@colorado.edu

Ryan Wills
University of Colorado Boulder
Ryan.wills@colorado.edu

PROBLEM STATEMENT / MOTIVATION

In 2016 alone, The Federal Trade Commission reported 3 million accounts of credit card fraud or identity theft related complaints. If companies can find trends in data and narrow down fraudulent activity, they will be able to find solutions as to why common areas may be targeting users. We strive to find complaints about fraudulent credit or debit card use and link these issues to specific Card Companies as well as finding trends in common zip codes.

REFERENCE:

Steele, J. 2017. Credit card fraud and ID theft statistics, digital. https://www.creditcards.com/credit-card-news/credit-card-security-id-theft-fraud-statistics-1276.php

1 LITERATURE SURVEY

Data in Action - Combatting Fraud: One company uses big-data analytics to find grey charges on user's credit cards and debit cards by drawing upon billing dispute data from the web, banks, and the CFPB's open consumer complaint database. 'Grey' charges are defined as lingering charges that a user previously signed up for a subscription or renewal service and may not be aware of the charge on their credit or debit card. While grey charges are not illegal, the user may not remember or have completely understood the terms presented and the charges can be misleading. Data.gov provides the highlight of a non-specific company combating grey charges using our same database, but the specifics of their analysis have not been released yet. Therefore, their research and techniques will not influence our Data Mining much, but it is important to note that work in this field is being conducted.

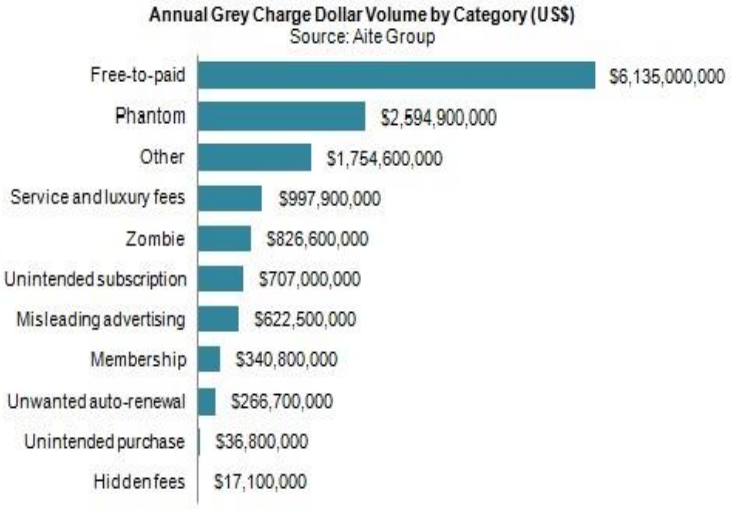

Figure 1: The Economic Impact of Grey Charges on Debit and Credit Card Issuers. 2012. Found at

https://aitegroup.com/report/economic-impact-grey-charges-debit-and-credit-card-issuers
Figure 1 clearly shows the Volume of Grey charges occurring in a single year – the front runner being free-to-paid subscriptions customers likely forget about. There is a fine line between fraud and grey charges and the described study works to understand the difference.

2 PROPOSED WORK

2.1 What Needed for Data Collection

It was fairly easy to collect our data. The Financial Services Consumer Complaint Database has been collecting this data for a long time and had a large data set that we were able to access. Since the data came from a .gov website, it was organized and formatted well and had plenty of data points.

2.2 Preprocessing

In order for our data to be usable, it will be necessary to clean up the large number of data points we have. Our data was organized well and due to it being collected from a good source, was formatted nicely. Inconsistent quotations will need to be synchronized and made consistent throughout the data set. There are many cases of empty cells, inconsistent formatting on cells, and partial zip codes. By getting rid of incomplete entries or making them zeros or empty strings, we should be able to make our data consistent, considering the data is in an excel type format. There are also several attributes that don't serve our purpose for the intent of this project that will also be removed.

2.3 How Mining Consumer Complaints is Different from Previous Work

For what we would like to do, we will not be looking at the types of charges that were made on the credit cards. We will be looking at the complaint, how the complaint was handles, the credit card company used, etc., without going into depth on purchases made with the credit cards. We wanted to avoid the purchases and financial part of the data, and focus on the customer service response from the customer complaint.

3 DATA SET

We are using the Financial Services Consumer Complaint Database found at: https://www.data.gov/consumer/. Currently downloaded on Cary's Mac (469 MB). Within the data set there are 18 attributes to include: Date received, Product, Sub-product, Issue, Sub-issue, Consumer complaint narrative, Company public response, Company, State ZIP code, Tags, Consumer consent provided, Submitted via, Date sent to company, Company response to consumer, Timely response, Consumer disputed, and Complaint ID. These attributes and data accumulate to over 17 million data points.

4 EVALUATION METHODS

Our results will be evaluated with cross validation. If we are able to use WEKA, there are built in evaluation tools to streamline the process. If we use Python, there are supporting libraries and methods to do the same.

5 TOOLS

We will be using Python and tools and libraries within such as mathplotlib, pandas, numpy, etc. Also, we hope to get familiar with WEKA in order to take advantage of its evaluation tools.

6 MILESTONES

- All data cleaned up and groomed – 19 March.
- Simple scatter plots and data correlation – 2 April.
- Using methods such as clustering and sequential patterns based on previous findings – 13 April.
- Analysis of all results – 19 April.
- Refactoring – 23 April.
- Final analysis and conclusion – 26 April.

7 SUMMARY OF PEER REVIEW SESSION

While we didn't receive much feedback, I believe we can work on a couple of things. First, our presentation must be more robust – we will have more slides and we will know our slides well before going up. This is a result of too little practice and confidence in the content. Secondly, our professionalism lacked. We should dress nicely, use notecards as quick references, speak clearly and loudly, and open up more to the audience.

A.1 Literature Survey

A.2 Proposed Work

A.2.1 What Needed for Data Collection

A.2.2 Preprocessing

Component Structures

Magnetization.

A.2.3 How Mining Consumer Complaints is different from previous work

A.2.4 Ground-State Magnetization Determination and DMM Micromagnetic Simulations

A.3 Data Set

A.4 Evaluation Methods

A.5 Tools

A.6 Milestones

A.7 Summary of Peer Review Sessions

ACKNOWLEDGMENTS

This work was partially supported by the MIUR-PRIN 2010–11 Project 2010ECA8P3 "DyNanoMag" and by the National Research Foundation, Prime Minister's office, Singapore under its Competitive Research Programme (CRP Award No. NRF-CRP 10-2012-03).

REFERENCES

[1] Patricia S. Abril and Robert Plant. 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan. 2007), 36–44. DOI: http://dx.doi.org/10.1145/1188913.1188915

[2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. 2002. Wireless Sensor Networks: A Survey. *Comm. ACM* 38, 4 (2002), 393–422.

[3] David A. Anisi. 2003. *Optimal Motion Control of a Ground Vehicle*. Master's thesis. Royal Institute of Technology (KTH), Stockholm, Sweden.

[4] P. Bahl, R. Chancre, and J. Dungeon. 2004. SSCH: Slotted Seeded Channel Hopping for Capacity Improvement in IEEE 802.11 Ad-Hoc Wireless Networks. In *Proceeding of the 10th International Conference on Mobile Computing and Networking* (MobiCom'04). ACM, New York, NY, 112–117.

[5] Kenneth L. Clarkson. 1985. *Algorithms for Closest-Point Problems (Computational Geometry)*. Ph.D. Dissertation. Stanford University, Palo Alto, CA. UMI Order Number: AAT 8506171.

[6] Jacques Cohen (Ed.). 1996. Special Issue: Digital Libraries. *Commun. ACM* 39, 11 (Nov. 1996).

[7] Bruce P. Douglass. 1998. Statecarts in use: structured analysis and object-orientation. In *Lectures on Embedded Systems*, Grzegorz Rozenberg and Frits W. Vaandrager (Eds.). Lecture Notes in Computer Science, Vol. 1494. Springer-Verlag, London, 368–394. DOI: http://dx.doi.org/10.1007/3-540-65193-429

[8] Ian Editor (Ed.). 2008. *The title of book two* (2nd. ed.). University of Chicago Press, Chicago, Chapter 100. DOI: http://dx.doi.org/10.1007/3-540-09237-4