Volume 11, Issue 1
March 2018
ISSN: 1946-1836

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

In this issue:

The **Journal of Information Systems Applied Research** (JISAR) is a double-blind peer-reviewed academic journal published by **ISCAP,** Information Systems and Computing Academic Professionals. Publishing frequency is three issues a year. The first date of publication was December 1, 2008.

JISAR is published online (http://jisar.org) in connection with CONISAR, the Conference on Information Systems Applied Research, which is also double-blind peer reviewed. Our sister publication, the Proceedings of CONISAR, features all papers, panels, workshops, and presentations from the conference. (http://conisar.org)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%), other journal papers (top 30%), unsettled papers, and non-journal papers. The unsettled papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in the JISAR journal. Currently the target acceptance rate for the journal is about 40%.

Questions should be addressed to the editor at editor@jisar.org or the publisher at publisher@jisar.org. Special thanks to members of AITP-EDSIG who perform the editorial and review processes for JISAR.

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

Editors

2018 JISAR Editorial Board

Information Sharing Increases Drug Sample Inventory Management Efficiency in Healthcare Clinics: Evidence from a Quasi-Experiment

Guido Lang
guido.lang@qu.edu
Quinnipiac University
Hamden, CT 06518

Ricky Chahal
ricky.chahal@checksamples.com
CheckSamples
New York, NY 10075

Abstract

Inefficient drug sample inventory management in healthcare clinics results in over $2.2 billion worth of drug samples being wasted in the United States every year. Pharmaceutical sales representatives are largely responsible for the forecasting, ordering, and delivery of drug samples in healthcare clinics. Thus, drug samples are a form of vendor-managed inventory, which requires inventory information sharing in order to be effective. A quasi-experimental study was conducted in order to assess the impact of information sharing on drug sample inventory management efficiency in healthcare clinics. A proprietary dataset of anonymized inventory transactions detailing the inflow and outflow of 19,400 drug samples, as well as the access data of said inventory information by pharmaceutical sales representatives was obtained from CheckSamples, a leading drug sample inventory management platform. Data collection took place during the nine month period from November 2016 to July 2017, covering multiple US-based clinics located in rural and urban settings, which range in size from single practitioners to clinics with over ten practitioners. Results indicate that information sharing improves inventory management efficiency, measured by average days in inventory, inventory days of supply, and dispense-through rate, by about 65% on average. Based on these results, information sharing in the context of drug samples holds the potential to generate significant cost savings while improving administrative efficiency and regulatory compliance. These findings are particularly relevant given the rising cost of healthcare and the associated policy debates in the United States today.

Keywords: information sharing, vendor-managed inventory, drug samples, healthcare

1. INTRODUCTION

Drug samples play a critical role in improving patient care by helping to establish preference, efficacy, and tolerance in patients, while reducing time to treatment and increasing drug adherence (Alikhan et al, 2010; Bastiaens, Chowdhury, & Gitelman, 2000). Moreover, drug samples provide access to medications among patients in high-risk groups (Tija et al., 2008). However, an examination of drug sample closets in healthcare clinics revealed that, on average 14% of medications were expired (Evans & Brown, 2012). Extrapolating this finding suggests that an estimated $2.2 billion worth of drug samples are wasted annually in the United States. This waste can be attributed to inefficient inventory management in healthcare clinics and should thus be preventable.

A closer examination of the drug sample inventory management process in healthcare

clinics points to the need for collaboration between pharmaceutical sales representatives and healthcare providers (Poser, 2007). Specifically, the responsibility for forecasting, ordering, and delivery of drug samples in healthcare clinics lies largely with pharmaceutical sales representatives. Thus, drug samples in healthcare clinics are an example of vendor-managed inventory (Hines et al, 2000).

Vendor-managed inventory (VMI) generally promises to lower inventory levels while increasing service levels (Levy and Grewel, 2000). However, in order to lead to such positive outcomes, VMI requires information sharing between buyer and vendor. In the context of drug sample inventory management, this suggests that drug sample inventory information should be shared between healthcare providers and pharmaceutical sales representatives in order to improve drug sample inventory management efficiency. Despite previous research on the topic of VMI in healthcare, the topic of drug sample inventory management and the benefits of information sharing in this context have been overlooked. The present study aims to address this gap by assessing the impact of information sharing on drug sample inventory management efficiency in healthcare clinics. The implications of this study are particularly relevant today, given the rapid growth of healthcare costs in the United States and the associated ongoing debate among policy-makers on how to combat this rise (e.g. Groves et al., 2013; Orszag & Ellis, 2007; Bodenheimer, 2005).

The remainder of this paper is structured as follows. The second section provides a brief overview of previous research on VMI implementations in the healthcare sector. The following section describes the methodology of the present study. Sections four and five present and discuss the results, while the last section summarizes this study's conclusions.

2. BACKGROUND

Vendor-managed inventory (VMI) has been defined as a collaborative initiative between a buyer and a vendor to optimize the availability of items and minimize cost to both network partners (Hines et al., 2000). Although VMI arrangements can take many forms (Christopher, 2016), the main goal of VMI is reducing inventory levels while improving service levels at the same time (Levy and Grewel, 2000). Performance benefits in VMI are generally achieved through information sharing between buyer and vendor and appropriate decision-making by the vendor (Sari, 2007). Information sharing, in particular of inventory information, is typically accomplished through information systems that provide real-time electronic data exchange (Yao & Dresner, 2008). Although VMI has been a popular topic in the logistics literature since the 1980s (Williams and Tokar, 2008), it has not received much attention in the healthcare sector until the early 2000s (Haavik, 2000). The following overview of recent studies examining VMI in the healthcare sector is meant to highlight the importance of, and difficulty associated with, implementing VMI in healthcare.

Enablers and performance outcomes associated with industrial vending systems in healthcare, which represent a specific form of VMI, were recently investigated by Falasca and Kros (2016). Their results suggest that the success of VMI in healthcare depends on the quality of the information generated by the information system and the quality of the buyer-vendor relationship. Moreover, their findings indicate that the successful implementation of VMI in healthcare can result in improved inventory management, enhanced service levels, and reduced costs.

An in-depth case study of VMI in a public, general multi-site hospital was conducted by Guimaraes and Carvalho (2013). They found that VMI led to significant improvements in inventory management, such as reduction of inventory costs, optimized inventory levels, decrease of emergency orders, and no stock-out episodes of pharmaceutical supplies. However, strong implementation barriers that are unique to the healthcare sector, such as regulation and a general lack of activity planning, were also found to hinder many of the benefits of VMI. The most significant factor in the successful implementation of VMI was found to be collaboration between partners and information sharing in the supply chain.

A survey of material managers and executives in healthcare by Callender and Grasman (2010) revealed that most respondents have received formal training and acquired appropriate skills and knowledge about supply chain best practices. However, an overwhelming majority of healthcare providers still believe that their inventory-related inefficiencies cannot be improved through information sharing and VMI. Although slightly outdated at this point, the findings of this study still point to a general need for better training and more education regarding the benefits of VMI in the healthcare sector.

Taken together, recent studies investigating VMI in the healthcare sector have generally found that VMI can cause significant improvements of inventory management efficiency. However, significant barriers, including regulation and education, appear to hinder the implementation of VMI in healthcare. The present study aims to contribute to the growing body of knowledge surrounding VMI in healthcare by focusing on the impact of information sharing on drug sample inventory management efficiency – a topic which has hitherto not been addressed in the literature.

3. METHODOLOGY

A proprietary dataset detailing the inventory of drug samples in multiple US-based healthcare clinics during the nine month period from November 2016 to July 2017 was obtained from CheckSamples. CheckSamples is a drug sample inventory management platform that helps healthcare providers increase administrative efficiency and ensure regulatory compliance with regards to the management and control of drug samples. In addition, CheckSamples provides pharmaceutical sales representatives the option to remotely access their clinics' drug sample inventory information, which allows them to optimize the supply of drug samples to clinics.

The dataset consists of anonymized inventory transactions detailing the inflow and outflow of all drug samples, as well as the access data of said inventory information by pharmaceutical sales representatives. The data stem from multiple US-based clinics, located in rural and urban settings, which range in size from single practitioners to clinics with over ten practitioners. During the nine month period from November 2016 to July 2017, a total of 19,400 drug samples were added to clinics' inventories, of which 8,954 (46.15%) were dispensed. The samples belong to 272 distinct drugs, which are made by 148 different pharmaceutical companies and represent 67 FDA Established Pharmacological Classes (EDCs). The top five EDCs are insulin analogs (25.86%), biguanides (6.49%), dipeptidyl peptidase 4 inhibitors (6.25%), l-thyroxines (5.60%), and GLP-1 receptor agonists (5.50%). For 214 (78.68%) of the 272 drugs, clinics' inventory information is not shared with any pharmaceutical representative. For the remaining 58 (21.32%) drugs, inventory information is actively shared with pharmaceutical sales representatives. Table 1 provides an overview of the dataset.

Three inventory management efficiency indicators were calculated for each drug: average days in inventory, inventory days of supply, and dispense-through rate. The indicators have been adapted to the context of drug sample inventory management based on established key performance indicators in inventory and supply chain management practice (Sylver, Pyke, & Thomas, 2017).

Table 1. Dataset Overview

Drugs and Drug Samples	
Drug samples added	19,400
Drug samples dispensed	8,954
Distinct drugs	272
Distinct pharmaceutical companies	148
Distinct FDA Established Pharmacological Classes	67
Information Sharing	
Drugs without information sharing	214
Drugs with information sharing	58

Average days in inventory (ADI) measures how long, on average, a sample of a particular drug is stored in inventory until it is dispensed. Healthcare providers and pharmaceutical sales representatives should strive to minimize it, since a shorter ADI indicates more efficient inventory management. The ADI is calculated as:

$$ADI_i = \frac{Total\ Days\ in\ Inventory_i}{Total\ Samples\ Dispensed_i}$$

where i denotes the drug. When making between-group comparisons, the ADI is averaged across all drugs.

Inventory days of supply (IDS) measures how long, on average, it would take to dispense the remaining sample inventory for a particular drug. Healthcare providers and pharmaceutical sales representatives should strive to minimize it, since a shorter IDS indicates more efficient inventory management. The IDS is calculated as:

$$IDS_i = \frac{Total\ Samples\ in\ Inventory_i}{Average\ Samples\ Dispensed\ per\ Day_i}$$

where i denotes the drug. Just like the ADI, the IDS is averaged across all drugs when making between-group comparisons.

The dispense-through rate (DTR) is a normalized measure of the amount of samples dispensed relative to the amount of samples added. Healthcare providers and pharmaceutical sales representatives should strive to maximize it,

since a larger DTR indicates more efficient inventory management. the DTR is calculated as:

$$DTR_i = \frac{Total\ Samples\ Dispensed_i}{Total\ Samples\ Added_i}$$

where i denotes the drug. Like the ADI and IDS, the DTR is averaged across all drugs when making between-group comparisons.

The data analysis exploits the access of clinics' drug sample inventory information by pharmaceutical sales representatives through the CheckSamples platform as an exogenous variable. This allows for between-group comparisons between drugs for which pharmaceutical sales representatives access clinics' drug sample inventory information (i.e. information sharing takes place) and drugs for which pharmaceutical sales representatives do not access clinics' drug sample inventory information (i.e. no information sharing takes place). Since pharmaceutical sales representatives are responsible for restocking clinics' drug sample inventories, one would expect better inventory management efficiency indicators under conditions of information sharing than under conditions of no information sharing. Thus, the research employs a single factor (no information sharing vs. information sharing) quasi-experimental design with three dependent variables (ADI, IDS, and DTR).

4. RESULTS

Average Days in Inventory

For the combined sample, average days in inventory (ADI) is about 35 days (M = 35.39, SD = 50.30). This suggests that, on average, drug samples remain in inventory for about 1.1 months before they are dispensed. Figure 1 depicts the difference in ADI between drugs with and without information sharing.

As depicted in Figure 1, for drugs without information sharing ADI is about 43 days (M = 42.75, SD = 53.98). In contrast, for drugs with information sharing, ADI is about 19 days (M = 18.55, SD = 35.70). Welch's t-test for difference in ADI between drugs without information sharing and drugs with information sharing is significant (t = 3.44, p < .001). In other words, drugs for which pharmaceutical sales representatives access clinics' drug sample inventory information remain in inventory for less than three weeks, whereas drugs for which pharmaceutical sales representatives do not access clinics' drug sample inventory information remain in inventory for over six weeks. Thus, ADI is significantly shorter (by 24 days, a decrease of 57%) for drugs with information sharing than for drugs without information sharing.

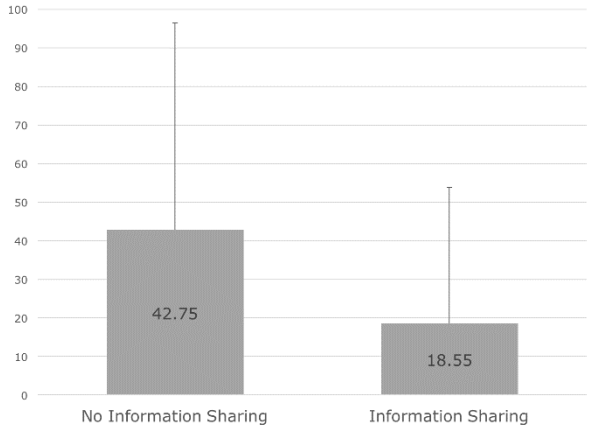


Figure 1: Average Days in Inventory

Inventory Days of Supply

For the combined sample, the inventory days of supply (IDS) is about 1,039 days (M = 1038.82, SD = 2134.95). This suggests that, on average, clinics' drug sample inventory lasts for about 2.8 years before being depleted. Figure 2 shows the IDS for drugs with and without information sharing.

Figure 2: Inventory Days of Supply

As shown in Figure 2, for drugs without information sharing the IDS is 1276 days (M = 1276.00, SD = 2457.71). In contrast, for drugs with information sharing, the IDS is about 496 days (M = 496.05, SD = 886.62). Welch's t-test for difference in IDS between drugs without information sharing and drugs with information sharing is significant (t = 3.04, p < .05). Stated differently, drugs for which pharmaceutical sales representatives access clinics' drug sample inventory information have inventories lasting

about 1.4 years, whereas drugs for which pharmaceutical sales representatives do not access clinics' drug sample inventory information have inventories lasting about 3.5 years. Thus, IDS is significantly shorter (by 780 days, a decrease of 61%) for drugs with information sharing than for drugs without information sharing.

Dispense-Through Rate

For the combined sample, the dispense-through rate (DTR) is 29% (*M* = 29.17%, *SD* = 35.32%). This suggests that, on average, less than one third of drug samples are dispensed within the study's nine month time frame. Figure 3 presents the DTR for drugs with and without information sharing.

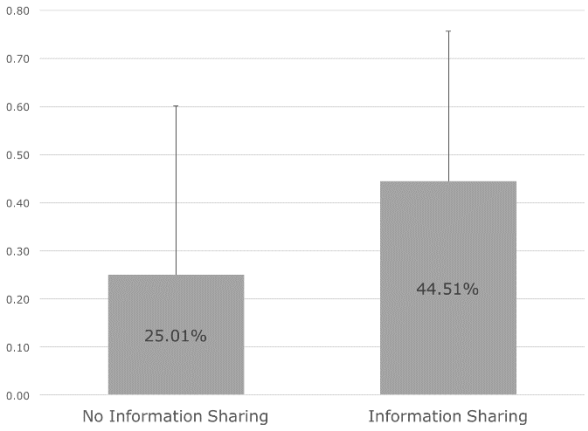


Figure 3: Dispense-Through Rate

As can be seen in Figure 3, for drugs without information sharing the DTR is 25% (*M* = 25.01%, *SD* = 35.23%). In contrast, for drugs with information sharing, the DTR is 45% (*M* = 44.51%, *SD* = 31.44%). Welch's t-test for difference in DTR between drugs without information sharing and drugs with information sharing is significant ($t = 3.97$, $p < .001$). In other words, almost half of the samples for drugs for which pharmaceutical sales representatives access clinics' drug sample inventory information are dispensed within nine months, whereas only a quarter of the samples for drugs for which pharmaceutical sales representatives do not access clinics' drug sample inventory information are dispensed within nine months. Thus, DTR is significantly larger (by 20%, an increase of 78%) for drugs with information sharing than for drugs without information sharing.

5. DISCUSSION

The inventory management of drugs for which pharmaceutical sales representatives access clinics' drug sample inventory information is significantly more efficient than that of drugs for which pharmaceutical sales representatives do not access clinics' drug sample inventory information. Specifically, the results indicate a 57% reduction in average days in inventory (ADI), a 61% reduction in inventory days of supply (IDS), and a 78% increase in dispense-through rate (DTR) for drugs with information sharing over drugs without information sharing. These figures suggest an average improvement of about 65% across all three indicators of inventory management efficiency. Given these results and the exogenous nature of information sharing in this quasi-experimental study, it seems plausible that information sharing causes improvements in inventory management efficiency.

Since inefficiencies in drug sample inventory management have been linked to over $2.2 billion worth of wasted samples per year, an improvement of about 65% could translate to over $1.4 billion worth of savings annually in the healthcare sector. Consequential benefits, such as improved administrative efficiency and regulatory compliance, are not even included in the figure of wasted samples. Thus, these findings indicate that information sharing holds tremendously valuable benefits for healthcare providers and pharmaceutical companies. A drug sample inventory management platform, such as CheckSamples, that allows for automated inventory information sharing between healthcare providers and pharmaceutical companies, is uniquely positioned to realize these benefits and create value for all parties involved.

However, the findings and implications of this study must be evaluated critically in light of its limitations. First, due to the quasi-experimental nature of this study, no random assignment of subjects to conditions took place. In particular, pharmaceutical sales representatives decided for themselves to access clinics' drug sample inventory information through the CheckSamples platform. Hence, it is possible that this self-selected sub-group differs from the overall group of pharmaceutical sales representatives with regards to attitudes and behaviors relevant to inventory management efficiency. Second, the three indicators of inventory management efficiency that are used in this study are not the only types of indicators that can be used to assess inventory management efficiency in the context of drug sample management. Thus, it is possible that the results of this study differ when other indicators for the dependent variables are employed. Third, the dataset, which consisted of

transactional drug sample inventory information from multiple US-based healthcare clinics during the time from November 2016 to July 2017 was limited in terms of its geographic scope, time span, and selection of clinics as well as pharmaceutical sales representatives. Therefore, it is possible that a sample with different characteristics will lead to different results with regards to the impact of information sharing on drug sample inventory management efficiency.

Future research may wish to explore several avenues to build on the foundation laid by this work. In particular, future research should try to design and implement a true randomized controlled study in which pharmaceutical sales representatives are randomly given the option to access clinics' drug sample inventory information. Moreover, future research should consider employing alternative indicators of inventory management efficiency, which may shed light on different aspects of inventory management efficiency in the context of drug samples. Lastly, future research would be well advised to expand the sample to a broader section of national and international healthcare clinics in order to mitigate the impact of potential selection bias.

6. CONCLUSION

Over $2.2 billion worth of drug samples expire ever year in the United States. This waste is due to inefficient drug sample inventory management practices in healthcare clinics. However, drug samples are vendor-managed, which means that pharmaceutical sales representatives are largely responsible for the forecasting, ordering, and delivery of drug samples in clinics. In situations of vendor-managed inventory, the sharing of inventory information with vendors is a crucial component of efficient inventory management.

A quasi-experiment was conducted in order to evaluate the impact of information sharing on inventory management efficiency in healthcare clinics. A proprietary dataset of anonymized inventory transactions detailing the inflow and outflow of 19,400 drug samples, as well as the access data of said inventory information by pharmaceutical sales representatives, across multiple US-based clinics during the nine month period from November 2016 to July 2017 was obtained from CheckSamples, a leading drug sample inventory management platform.

Results indicate that information sharing improves inventory management efficiency by about 65%. Based on the value of wasted drug samples, information sharing between healthcare providers and pharmaceutical companies could lead to savings of about $1.4 billion annually. These savings, along with consequential benefits of improved administrative efficiency and regulatory compliance, appear to be particularly attractive given the growth of healthcare costs and associated policy debates in the United States.

7. REFERENCES

Alikhan, A., Sockolov, M., Brodell, R. T., & Feldman, S. R. (2010). Drug Samples in Dermatology: Special Considerations and Recommendations for the Future. *Journal of the American Academy of Dermatology*, *62*(6), 1053-1061.

Bastiaens, L., Chowdhury, S., & Gitelman, L. (2000). Medication Samples and Drug Compliance. *Psychiatric Services, 51*(6), 819-819.

Bodenheimer, T. (2005). High and Rising Health Care Costs. Part 1: Seeking an Explanation. *Annals of Internal Medicine, 142*(10), 847-854.

Callender, C. & Grasman, S. E. (2010). Barriers and Best Practices for Material Management in the Healthcare Sector. *Engineering Management Journal, 22*(4), 11-19.

Christopher, M. (2016). *Logistics & Supply Chain Management, Fifth Edition*. FT Press: Upper Salle River, NJ.

Evans, K. L., & Brown, S. R. (2012). Many Sample Closet Medications Are Expired. *Journal of the American Board of Family Medicine, 25*(3), 394-395.

Falasca, M., & Kros, J. F. (2016). Success Factors and Performance Outcomes of Healthcare Industrial Vending Systems: An Empirical Analysis. *Technological Forecasting & Social Change*, in press.

Groves, P., Kayyali, B., Knott, D., & Van Kuiken, S. (2013). *The 'Big Data' Revolution in Healthcare: Accelerating Value and Innovation*. Report of the Center for US Health System Reform, Business Technology Office. McKinsey & Company: New York, NY.

Guimaraes, C. M. & Carvalho, J. C. (2013). Vendor Managed Inventory (VMI): Evidences from Lean Deployment in Healthcare.

Strategic Outsourcing: An International Journal, 6(1), 8-24.

Haavik, S. (2000). Building a Demand-Driven, Vendor-Managed Supply Chain. *Healthcare Financial Management, 24*(2), 56-61.

Hines, P., Lamming, R., Jones, D., Cousins, P., & Rich, N. (2000). *Value Stream Management: Strategy and Excellence in the Supply Chain*. Financial Times Prentice Hall, Upper Saddle River, NJ.

Levy, M. & Grewel, D. (2000). Supply Chain Management in a Networked Economy. *Journal of Retailing, 76*(4), 415-29.

Orszag, P. R. & Ellis, P. (2007). Addressing Rising Health Care Costs – A View from the Congressional Budget Office. *The New England Journal of Medicine, 357*, 1885-1887.

Poser, S. (2007). Unlabeled Prescription Drug Samples and the 'Learned Intermediary': The Case for Liability without Preemption. *Food and Drug Law Journal, 62*, 653–694.

Sari, K. (2007). Exploring the Benefits of Vendor Managed Inventory. *International Journal of Physical Distribution & Logistics Management, 37*(7), 529-545.

Sylver, E. A., Pyke, D. F., & Thomas, D. J. (2017). *Inventory and Production Management in Supply Chains, Fourth Edition*. CRC Press: Boca Raton, FL.

Tija, J., Briesacher, B. A., Soumerai, S. B., Pierre-Jacques, M., Zhang, F., Ross-Degnan, D., & Gurwitz, J. H. (2008). Medicare Beneficiaries and Free Prescription Drug Samples: A National Survey. *Journal of General Internal Medicine, 23*(6), 709-714.

Williams, B.D. & Tokar, T. (2008). A Review of Inventory Management Research in Major Logistics Journals. *The International Journal of Logistics Management, 19*(2), 212-32.

Yao, Y. & Dresner, M. (2008). The Inventory Value of Information Sharing, Continuous Replenishment, and Vendor-Managed Inventory. *Transportation Research Part E: Logistics and Transportation Review, 44*(3), 361-378.

A Dynamic and Static Analysis of the Uber Mobile Application from a Privacy Perspective

Darren R. Hayes
dhayes@pace.edu

Christopher Snow
csnow@pace.edu

Saleh Altuwayjiri
sa07549n@pace.edu

Pace University
New York, NY

Abstract

This research examined the Uber mobile application and determined that the company utilizes very precise and potentially invasive geolocation tracking techniques. Our experiments indicate that Uber tracks the location of its users, after the conclusion of a ride, for longer than its official privacy policy would indicate. The most interesting finding is that geolocation tracking is performed by the Uber app even when the user does not take an Uber ride. Additionally, geolocation tracking of users using competing services may be disconcerting for some consumers. While our findings may be a privacy concern for some consumers, the Uber app also has tremendous potential for digital forensics investigators.

Keywords: Uber, mobile apps, mobile forensics, geolocation, privacy

1. INTRODUCTION

Uber is a service that enables drivers to act as flexible contractors and provide transportation services that compete with traditional taxi services. Consumers, using the Uber mobile app, can search for a car service in their area. The benefit to the consumer is that they are visually provided with the mapped location of Uber cars in their vicinity and are provided with an upfront quote for a specific journey (or "ride"). Uber operates in 570 cities worldwide. In recent times, Uber has received negative press about its geolocation tracking of users, which raised a number of concerns regarding its privacy policies and potentially invasive data collection practices. The research, herein, sought to identify the type of personally identifiable information (PII) collected by the company. More precisely, this research identifies what geolocation data the Uber mobile application (app) collects on the user.

In April 2017, the New York Times published a story that documented a meeting, at Apple headquarters, in 2015, between Travis Kalanick, CEO of Uber, and Tim Cook, CEO of Apple (Issac, 2017). The article alleges that Mr. Cook scolded Mr. Kalanick for identifying and tagging iPhones after the Uber app had been uninstalled or the device had been wiped. Apparently, this type of user identity coding violated the Apple developer terms of service agreement (Newman, 2017).

An article in The New York Times detailed how Unroll.me, which purported to purge your device's email inbox of annoying advertising messages, was actually being used to spy on competitors (Isaac & Lohr, 2017). The article documented how Unroll.me would scan a user's

inbox, identify if there were service receipts, from competing companies like Lyft, and then sell that information to Lyft's competitor - Uber.

2. BACKGROUND

Since the introduction of iOS 5, Apple has been limiting app developer access to the iPhone's UDID (Unique Device Identifier)(Schonfeld, 2011). A notice from Apple stated, "Starting May 1, the App Store will no longer accept new apps or app updates that access the UDID; please update your apps and servers to associate users with the Vendor or Advertising identifiers introduced in iOS 6" (Panzarino, M.). Apple now prefers that app developers utilize the official Apple advertising platform to track app users. Based on Apple's *Advertising and Privacy* policy, it appears that Apple does collect user data and then subsequently shares it with third-parties (Apple, 2017). Nevertheless, developers can obtain extensive information about an app user through the integration of the *UIDevice* object. The UIDevice object can be used by an app developer to determine the assigned name of the device, device model and iOS version, orientation (*orientation* property) of the device, battery charge (*batteryState* property) and distance of the device to the user (*proximityState* property) (Apple). Moreover, developers can integrate code, during app development, for third-party analytics. These third-party companies include Localytics, mixpanel, UXCam and fabric. Companies like Apptopia provide app developers with extensive, nay invasive, analytics on competitor apps.

The use of the user UDID has not always been used for nefarious purposes, however; the UDID was often used to identify if the user of an app was a legitimate user and could block a user's access if an account was compromised or potentially stolen. Fingerprinting is yet another methodology, used by third-parties, to uniquely identify users, based on application configuration. Fingerprinting is best known for identifying online users based on user settings from their browser, which may include user cookies and browser plug-ins. The Electronic Frontier Foundation (EFF) created a project known as Panopticlick (panopticlick.eff.org) to raise awareness about how your browser is used by advertisers, and others, to identify and track you on the Web. The EFF announced that 84% of online users can be uniquely identified by their browser (Budington, 2015).

According to Uber's *USER PRIVACY STATEMENT*, there are two categories of information collected about users: (a) Information You Provide to Us, which can include name, email, phone number, postal address, profile picture, payment method, and (b) Information We Collect Through Your Use of Our Services, which can include location information, contacts, transactions, usage and preference, device information, call and SMS data and log information (Uber, 2015). Of particular interest is the Device Information (hardware model, operating system and version, software and file names and versions, preferred language, unique device identifier, advertising identifiers, serial number, device motion information, and mobile network information). In terms of location information, Uber is not specific about the extent to which the user's location is being tracked but states that they "may also collect the precise location of your device when the app is running in the foreground or background" (Uber, 2015). Uber provides more detailed information about the use of Location Services on its Website under iOS App Permissions (Uber).

What is most interesting, for the purposes of our research, is that during our installation of the Uber app, a dialog box appears and states that "Uber collects your location (i) when the app is open and (ii) from the time of the trip request through five minutes after the trip ends" (see Figure 1). Uber states in their FAQ that the reasoning behind this data collection is to "improve pickups, drop-offs, customer service, and to enhance safety" (Uber). However, users reported seeing the Uber app using location services weeks after the app was used and certainly beyond the stated 5 minutes. Uber responded to these reports blaming Apple's iOS Maps extension that Uber uses to serve regional maps to their customers (Perez, 2016).

Perhaps unsurprisingly, Uber has invested heavily in data science to retain its competitive advantage, as evidenced by its aggressive recruitment of data scientists (Wilde, 2015). We also know that Uber extensively uses a telematics pilot program, called *Autohawk*, to identify the location of its drivers and perform diagnostic testing on the vehicle to ensure passenger safety (Wisniewski, 2016). In fact, Uber provides geolocation information, provided by its data visualization team, on its Website at eng.uber.com/data-viz-intel. Uber integrates both Fabric and Localytics in its mobile app. Fabric provides companies, like Uber, with real-time information about the health of their app. These analytics include application crash analytics.

As of November 2017, allegations abound about Uber's competitor spy programs. The *Waymo v.*

Uber lawsuit appears to indicate that Uber may have been involved in illegal espionage. A letter, submitted as evidence in this lawsuit and penned by Richard Jacobs, former Uber security executive, details Uber's illegal practices of hiring actors to collect data and spy on their competitors. In the letter, Jacobs, who at the time had filed suit against Uber in the capacity of "whistleblower," details practices that would lead to the theft of tradesecrets related to competitor fares and driver incentives. To settle, Uber paid Jacobs $4.3 million at the time. His allegations have now been made public and have been used in a related case involving the self-driving car service Waymo. In this case, a former employee allegedly sold tradesecrets to Uber prior to the company being acquired by Uber (Marshall, 2017).

3. RELATED WORK

The research group at Pace University has previously published research on how geolocation information from mobile apps could be used by governments to track users in an article entitled "Leakage of Geolocation Data by Mobile Ad Networks" (Snow, Hayes, Dwyer, 2016) and another article entitled "Mobile Ad Networks and Security Issues Regarding Geolocation Data" (Snow, Hayes, Dwyer, 2016). More specifically, this research focused on the claims of the whistleblower, Edward Snowden, that mobile apps, like Angry Birds, could be used to profile individuals and track their whereabouts through communications with mobile advertisements.

EURCOM's research paper, entitled "Taming the Android AppStore: Lightweight Characterization of Android Application", examined the network connections established by "legitimate" apps (Vigneri, Chandrashekar, Pefkianakis, & Heen, 2015). Ultimately, their research determined that the 2,146 applications that they examined connected to approximately 250,000 distinct URLs; many of these server connections were to domains known to contain malware.

A recent paper published by IEEE titled "Location Privacy Breach: Apps Are Watching You in Background" (Liu, D., Gao X., & Wang, H.) examined the correlation between the frequency of locational requests and potential security risks associated with user data leakage. The study concluded that shorter intervals between each geolocation request from a mobile app correlate to a higher privacy risk. Therefore, mobile applications that constantly update with a user's location, such as a car service are at high risk of leaking a user's location because of their high request rate.

It is also known that mobile users are not generally aware when an app is accessing personal data in the background, as noted in the research study "Android permissions remystified: a field study on contextual integrity" published by USENIX Security (Wijesekera P., Baokar A., Hosseini A., Egelman S., Wagner D., and Beznosov K.). The 2015 study analyzed how 36 participants used their mobile applications and concluded that 80% of participants studied noted that at least one permission request, made by a mobile application, as inappropriate. These permission requests were approved when the user accepted the end-user agreement during installation. The participants were unaware about exactly what they were allowing the mobile application to access. Thus, a majority of the participants experienced an invasion of (locational) privacy from the mobile applications they used.

4. EXPERIMENTAL RESEARCH

Our methodology, for this research, involved performing both a static and dynamic analysis of the Uber mobile application for iOS (iPhone) and for Android. The static analysis primarily involved an analysis of the SQLite database associated with the app, utilizing mobile forensic tools on an iPhone. Additionally, our static analysis included reverse engineering the Android application package (APK) file. The purpose of the latter was to review the code to identify the application's manifest. This app manifest would include permissions on the user device, which could include access to contacts, user location (based on GPS, cell sites or local access points), access to device hardware, like the camera or microphone and identifying information about the user. The dynamic analysis focused on how the application behaved during execution. The latter would use a series of DNS analytical tools to identify domain and network connections.

Google Maps

The Uber app for Android utilizes the Google Maps API (application programming interface) (see Figure 2) for locating Uber drivers in the vicinity of the consumer. We also confirmed this through our dynamic analysis of PCAPs (packet capture) from the Uber app (see Figure 3).

APK Analysis

An examination of the Uber APK file provides a list of permissions requested by the application. A review of the APK manifest quickly identifies the

location permissions requested for the user. A search for the "Uber APK file" quickly identifies where the application can be downloaded from the Web. Once downloaded there are a number of applications that can be used to review the code and manifest for the APK. One tool for reviewing APK developer code is *dex2jar* (dex compiler), which can be downloaded from SourceForge. Yet another application for viewing the APK is FileViewer Plus. For our research we chose to use an online APK decompiler application, which is available from www.javadecompilers.com/apk. The rationale for selecting this tool to decompile the APK was that no download was required and the APK file can simply be uploaded on the fly. An analysis of the APK manifest reveals the following permissions related to location:

android.permission.ACCESS_COARSE_LOCATION
android.permission.ACCESS_FINE_LOCATION

There is nothing unusual about these location permissions being requested by the Uber app. The location of a user device can be based on GPS, proximity to a cellular tower (cell site) or WiFi. ACCESS_COARSE_LOCATION is a permission that enables the app to access the approximate location of the user device, which is based on NETWORK_PROVIDER (cell sites). ACCESS_FINE_LOCATION enables the app to determine the location of the user device based on NETWORK_PROVIDER and GPS (GPS_PROVIDER).

Using BlackLight (BlackBag Tech) and MPE+ (AccessData) forensics tools, we can determine that the data associated with the Uber app can be found in the following application files:

Library>Application Support>com.ubercab.UberClient\xxxx.ldb

Library>Application Support>com.ubercab.UberClient\Cache.db

Data>data>com.ubercab\files\rider\0000xx.ldb

Analysis Background
The data below was compiled using two analyst programs:

- Debookee by iwaxx
- BlackLight by BlackBag

The initial experiments were run from an iPhone 7 Plus device running iOS 10.2.1. The Uber app was downloaded for the first time on this new iPhone. The user account on the Uber app was new and no ride history existed. Actions that were performed include searching for a nearby Uber car, mapping out a pick-up point and destination address to determine ETA (estimated time of arrival) and price in addition to adding a PayPal payment method.

Man-in-the-Middle Analysis
The highlighted lines in Figure 4 are HTTPS requests made by the Uber app. These lines were captured using Debookee, a LAN scanning program by iwaxx; the tool can capture traffic on a targeted device. Each request uses the SSL, under the HTTPS protocol, as seen in the "method" column. The HTTPS protocol ensures that data being pushed from the user, like location, to the Uber servers is encrypted and therefore illegible to bad actors listening to the connection. One can infer that *cn-geo1.uber.com* is a domain used specifically for transferring information about a user's location and nearby Uber vehicles.

Location Analysis with BlackLight
Using BlackLight, we were able to find an interesting file called "eyeball", as seen in Figure 5.

"etaString":"5 minutes","etaStringShort":"5mins","averageEta":292,"minEta":5},"39":{" vehiclepaths":{"8d8269c664cbd9342acbdd14b5cf5ccca82fd2d4":[{"latitude":40.70981, "course":133,"longitude":-74.00881,"epoch":1497472606280},{"latitude":40.70974, "course":133,"longitude":-74.00870999999999,"epoch":1497472610512}, {"latitude":40.70969,"course":133,"longitude":-74.00864,"epoch":1497472614714}, {"latitude":40.70959000000001,"course":133,"longitude":-74.00851,…

The above string is a sampling of many similar strings that were found in a file labeled "eyeball" under the following folder pathway in the UberClient application folder:

Library>Application Support>PersistentStorage>BootstrapStore>RealtimeApp.StreamModelKey

It can be inferred that "eyeball" refers to the radius circle of a user's location to capture nearby Uber vehicles. This "etaString" operation is repeated throughout the "eyeball" storage file for multiple vehicles in the area. The string explains an Estimated Time of Arrival (ETA) of 5 minutes

and provides an array of latitude and longitude numbers connected to a "vehicle path" followed by a string of numbers and letters that can be assumed to be a serial number for the specific vehicle. When the various longitude and latitude coordinates in the array are mapped out, the points display a route that this specific nearby Uber vehicle was following. Figure 6 shows the latitude and longitude coordinates mapped out on Google Maps. It can be inferred that each of these coordinates was a ping from the nearby Uber vehicle to the app in order to update the driver's position on the user's screen.

The most important string from this same "eyeball" storage file came at the very end. The following string shows a "reverseGeocode" operation that was able to identify the exact address of where the user was located.

"reverseGeocode":{"latitude":40.7102306652868,"components":[{"long_name":"163 William Street","short_name":"163 William Street","types":["premise"]},{"long_name":"Lower Manhattan","short_name":"Lower Manhattan","types":["neighborhood","political"]},{"long_name":"Manhattan","short_name":"Manhattan","types":["political","sublocality","sublocality_level_1"]},{"long_name":"New York","short_name":"New York","types":["locality","political"]},{"long_name":"New York County","short_name":"New York County","types":["administrative_area_level_2","political"] (…) **"longAddress":"163 William Street, New York, NY 10038, USA","nickname":"163 William Street","uuid":"3938134a-1b86-4f87-8ad0-f29c66ea674d","longitude":-74.00613835924165,"shortAddress":"163 William Street"}}**

This "reverseGeocode" protocol was able to accurately identify the user's current location as "163 William Street" without this exact address being explicitly provided. The application then used this geocode to find Uber vehicles in the area. We can therefore deduce that a user's location is saved as well as the vehicles in the area, regardless of whether an Uber is requested for a pick-up. Even if the "reverseGeocode" was not provided, the previous "vehiclepaths" could be mapped out to determine the general location of where the Uber customer stood by seeing where nearby Uber vehicles were pinging.

Privacy Analysis with BlackLight

[{"useCase":"personal","hasBalance":false,"status":"active","accountName":"Apple Pay Display","tokenDisplayName":"Apple Pay Display","tokenType":"apple_pay_display","uuid":"f8b461f2-a12e-4e3f-afdb-682c79497726"},{"useCase":"personal","cardNumber":"payp","cardExpiration":"20-**-12T14:48:11.257+00:00","cardType":"PayPal","hasBalance":false,"status":"active","accountName":"*********@me.com","cardExpirationEpoch":1812811691257,"tokenDisplayName":"*********@me.com","uuid":"3a48e583-71dc-4a51-82f0-01ddd8b7106b","tokenType":"paypal"}]**

**portions redacted for privacy*

The above string was found in a file labeled "profiles" under the following folder pathway in the UberClient application folder:

Library>Application Support>PersistentStorage>Store>PaymentBase.PaymentStreamModelKey

The string displays two payment methods that were saved on the Uber app; the first being Apple Pay and this method was not set up on the target device; the second being PayPal, which was the connected method of payment when requesting rides. The redacted portions represent areas where the user's PayPal login username was saved in plaintext. Other fields for "cardExpiration as well as "cardNumber" are also visible.

As part of our static analysis, we analyzed searches for trips and actual trips using the Uber app on an iPhone. Using BlackLight, we determined that all of the locations of pick-ups and drops offs were found in a file called database.db located here:

Root>Mobile>Applications>UberClient>database.db

In the first experiment, the Uber user searched for an address in Brooklyn (3xx Gold St, Brooklyn, New York 11201). The user decided to take a NYC Yellow Cab instead of using Uber, yet the app tracked the user's trip with a competitor service, as noted in Figure 7. Interestingly, using BlackLight, we could determine that the Yellow Cab journey was recorded in a very similar way as an Uber journey, as shown in Figure 8.

In the second experiment, the user used the Uber app to find the cost of a journey. The Uber app

was then closed. The user then used the Lyft app (competitor service) to see if the journey was cheaper than Uber. Given the cost saving, the user selected the ride with the Lyft app. The user took a Lyft ride from the Cheesecake Factory (100 Cambridge Side Place, Cambridge, MA 02141) to Harvard University. Using BlackLight, we discovered that this Lyft ride was recorded in the Uber app, as shown in Figures 9 and 10. Furthermore, we found that the Uber app was tracking the user for approximately 11 minutes after the Lyft ride had ended, thereby negating Uber's claim that a user will only be tracked for up to 5 minutes after the conclusion of a ride, although the ride was with a competing service. We proved this by reviewing the route, recorded by the app's Google API, after the ride had ended and determined that it took the user 11 minutes to complete the mapped route.

5. FUTURE EXPERIMENTS

Our research discovered that the Uber app uses Crashlytics. This third-party analytics service, which was acquired by Google, provides the app developer with crash analytics. Theoretically, while this service is providing real-time crash analytics, it could also be collecting the location of the user device. During our dynamic app analysis with Debookee, we discovered that there were continuous HTTPS requests going back and forth between the device and the Crashlytics server(s), even though the app was not experiencing any crash problems. Additionally, a more extensive examination of comparable "ride" apps would be appropriate to identify whether there are simply privacy concerns with Uber and not with its competitors.

The new iOS 11 update provides enhanced controls for the user by forcing apps, old and new, to prompt the user to select whether or not they want the application to share their location always, only when the app is being used, or never allow location sharing. An investigation to see if any transportation apps, utilizing the location permission, are somehow bypassing a user's preference is certainly of interest (Fleishman, 2017).

Moreover, as of August 29, 2017, the controversial practice that allowed Uber to continue mobile tracking post-ride was removed - a change confirmed by Uber's Chief Security Officer, Joe Sullivan. Future experiments should seek to identify if this invasive tracking has been effectively removed or if post-ride location tracking, in addition to competitor tracking, has in fact been halted (Volz, 2017).

6. CONCLUSIONS

In an age where many people, especially in the United States, are concerned about government agencies, like the NSA, collecting vast quantities of PII and geolocation information, consumers should also understand the data collection practices of companies, including Uber. Even though Uber placed blame for their app accessing locational services after 5 minutes on Apple's iOS Maps, it is clear from our findings that their application could store locational data after 5 minutes. With the application able to access location from the iOS Map issue that they have cited, the Uber app may still then be able to tap into a user's location past the 5 minutes indicated in their privacy agreement. The Uber mobile application is saving this location data, from vehicles in a user's area, locally on the user's device. In addition, the user's exact address locations are being determined and saved on the device through longitude and latitude coordinates. It is clear that Uber is not just saving trip locations from completed rides but are collecting geolocation data when the app is not being used for a ride and, more interestingly, this information is being used to monitor rides with competing services. This coincides with new evidence arising from the *Waymo v. Uber* case, where it has become aparent that Uber is not opposed to spying on its competitors. From a forensics perspective, this research demonstrates the tremendous potential for criminal investigators to use the Uber app to track a suspect and/or victim. The value of this information cannot be overstated given the detailed unencrypted user information available, in plaintext, using forensic imaging tools, like BlackLight.

7. REFERENCES

Apple (2017, January 16). About Advertising and Privacy. Retrieved June 10, 2017, from https://support.apple.com/en-us/HT205223

Apple. UIDevice: A representation of the current device. Retrieved June 10, 2017, from https://developer.apple.com/documentation/uikit/uidevice

Budington, Bill (2015, December 17). Panopticlick 2.0 Launches, Featuring New Tracker Protection and Fingerprinting Tests. Retrieved June 11, 2017, from https://www.eff.org/deeplinks/2015/12/panopticlick-20-launches-featuring-new-tracker-protection-and-fingerprinting-tests

Feishman, Glenn (2017, July 10). How iOS 11 changes locaton tracking on your iPhone and iPad. Retrieved December 20, 2017 from https://www.macworld.com/article/3203365/ios/how-ios-11-changes-location-tracking-on-your-iphone-and-ipad.html

Liu, D., Gao X., & Wang, H. (2017, July 17). Location Privacy Breach: Apps Are Watching You in Background, IEEE 37th International Conference on Distributed Computing Systems (ICDCS), 2017.

Google. (2015, April 17). Inside AdWords: ads take a step towards "HTTPS everywhere". Retrieved June 10, 2017, from http://adwords.blogspot.com/2015/04/ads-take-step-towards-https-everywhere.html

Hof, R. (2014, August 27). Study: mobile ads actually do work - especially in apps. Retrieved June 10, 2017 from http://www.forbes.com/sites/roberthof/2014/08/27/study-mobile-ads-actually-do-work-especially-in-apps/

Isaac, Mike (2017, April 23). Uber's C.E.O. Plays With Fire. Retrieved June 10, 2017 from https://www.nytimes.com/2017/04/23/technology/travis-kalanick-pushes-uber-and-himself-to-the-precipice.html?_r=1

Isaac, Mike, & Lohr, S. (2017, April 24). Unroll.me Service Faces Backlash Over a Widespread Practice: Selling User Data. Retrieved June 10, 2017 from https://www.nytimes.com/2017/04/24/technology/personal-data-firm-slice-unroll-me-backlash-uber.html

Lee, M. (2015, January 26). Secret 'BADASS' intelligence program spied on smartphones. *First Look Media*. Retrieved June 10, 2017, from https://firstlook.org/theintercept/2015/01/26/secret-badass-spy-program/

Marshall, Aarian (2017, December 20). The 37-Page Letter That Could Make Uber's Problems So Much Worse. Retrieved December 20, 2017 from https://www.wired.com/story/uber-waymo-jacobs-letter/

Newman, Lily Hay (2017, April 24). Uber Didn't Track Users Who Deleted the App, But it Still Broke the Rules. Retrieved June 10, 2017 from https://www.wired.com/2017/04/uber-didnt-track-users-deleted-app-still-broke-rules/

Panzarino, Matthew (2013, March 21). Apple to reject any apps that use UDIDs, don't support Retina, iPhone 5 displays as of May 1st. Retrieved June 10, 2017 from https://thenextweb.com/apple/2013/03/21/after-a-year-of-warnings-apple-will-no-longer-accept-any-apps-that-use-udids-as-of-may-1st/

Perez, Sarah (2016, Dec 22). Uber Explains Why It Looks like Its App Is Still Tracking Your Location, Long after Drop-Off, Tech Crunch. Retrieved August 1, 2018 from https://techcrunch.com/2016/12/22/uber-explains-why-it-looks-like-its-app-is-still-tracking-your-location-long-after-drop-off/

Schonfeld, Erick (August 19, 2011). Apple Sneaks A Big Change Into iOS 5: Phasing Out Developer Access To The UDID. Retrieved June 10, 2017, from https://techcrunch.com/2011/08/19/apple-ios-5-phasing-out-udid/

Snow, C., Hayes, & D., Dwyer, C. (2016). Leakage of Geolocation Data by Mobile Ad Networks, Journal of Information Systems Applied Research, 2016.

Snow, C., Hayes, & D., Dwyer, C. (2015). Mobile Ad Networks and Security Issues Regarding Geolocation Data, EDSIG Conference 2015.

Uber. iOS App Permissions. Retrieved June 11, 2017 from https://www.uber.com/legal/other/ios-permissions/

Uber (2015, July 15). User Privacy Statement. Retrieved June 11, 2017 from https://www.uber.com/legal/privacy/users/en/

Uber.). How does Uber use location (Android)? Retrieved July 31, 2017 from https://help.uber.com/h/ba9dd342-158d-421f-a9ea-0e6c7aaad726

Vigneri, L., Chandrashekar, J., Pefkianakis, I., & Heen, O. (2015). Taming the Android AppStore: lightweight characterization of Android applications. *arXiv preprint arXiv:1504.0609*

Volz, Dustin (2017, August 29). Uber to end post-trip tracking of riders as privacy push. Retreived December 20, 2017 from https://www.reuters.com/article/us-uber-privacy/uber-to-end-post-trip-tracking-of-riders-as-part-of-privacy-push-idUSKCN1B90EN

Wijesekera P., Baokar A., Hosseini A., Egelman S., Wagner D., and Beznosov K. (2015). Android permissions remystified: A field study

on contextual integrity, Proceedings of the 24th USENIX Conference on Security Symposium

Wilde, Ben (2015). Data Science Disruptors: How Uber Uses Applied Analytics For Competitive Advantage. Retrieved June 11, 2017 from https://georgianpartners.com/data-science-disruptors-uber-uses-applied-analytics-competitive-advantage/

Wisniewski, Mary (2016). Uber says monitoring drivers improves safety, but drivers have mixed views. Retrieved June 11, 2017 from http://www.chicagotribune.com/news/local/breaking/ct-uber-telematics-getting-around-20161218-column.html

Editor's Note:

This paper was selected for inclusion in the journal as the CONISAR 2017 Best Paper The acceptance rate is typically 2% for this category of paper based on blind reviews from six or more peers including three or more former best papers authors who did not submit a paper in 2017.

Appendices and Annexures

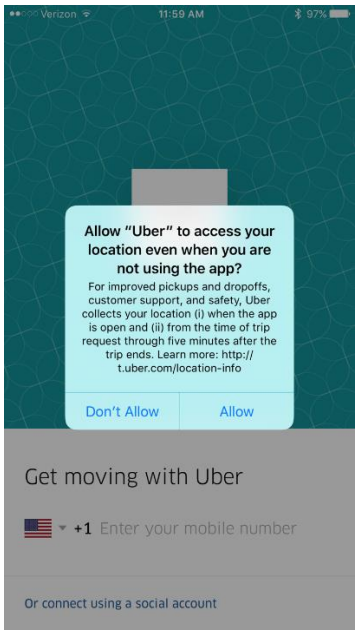


Figure 1. Uber dialog box during installation

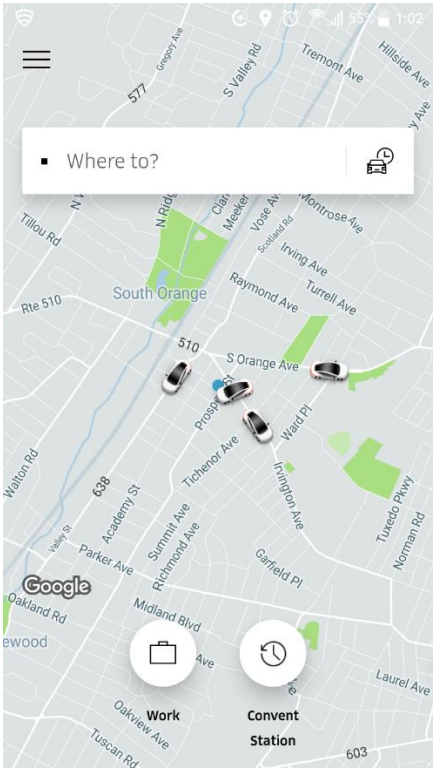


Figure 2. Screenshot of Google Maps in the Uber app

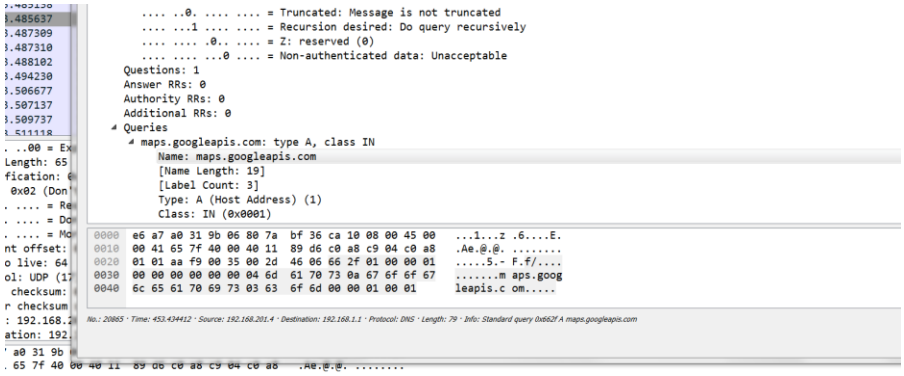


Figure 3. Google Maps API identified in a PCAP captured by Wireshark

Time	URL	Method
20:57:28	https://settings.crashlytics.com/...[encrypted]	HTTPS
20:57:28	https://e.crashlytics.com/...[encrypted]	HTTPS
20:57:28	https://blackswan.uber.com/...[encrypted]	HTTPS
20:57:30	https://cn-dca1.uber.com/...[encrypted]	HTTPS
20:57:30	https://cl3.apple.com/...[encrypted]	HTTPS
20:57:30	https://cl3.apple.com/...[encrypted]	HTTPS
20:57:30	https://static.gc.apple.com/...[encrypted]	HTTPS
20:57:30	https://cn-dca1.uber.com/...[encrypted]	HTTPS
20:57:31	https://clients4.google.com/...[encrypted]	HTTPS
20:57:32	https://d1a3f4spazzrp4.cloudfront.net/...[encrypted]	HTTPS
20:57:32	https://cn-geo1.uber.com/...[encrypted]	HTTPS
20:57:32	https://service.gc.apple.com/...[encrypted]	HTTPS
20:57:33	https://service.gc.apple.com/...[encrypted]	HTTPS
20:57:33	https://service.gc.apple.com/...[encrypted]	HTTPS
20:57:33	https://service.gc.apple.com/...[encrypted]	HTTPS
20:57:33	https://init.itunes.apple.com/...[encrypted]	HTTPS
20:57:34	https://play.itunes.apple.com/...[encrypted]	HTTPS
20:57:34	https://xp.apple.com/...[encrypted]	HTTPS
20:57:34	https://sp.itunes.apple.com/...[encrypted]	HTTPS
20:57:39	https://configuration.apple.com/...[encrypted]	HTTPS
20:58:00	https://cn-geo1.uber.com/...[encrypted]	HTTPS
20:58:24	https://cn-geo1.uber.com/...[encrypted]	HTTPS

Figure 4. HTTPS requests initiated by Uber

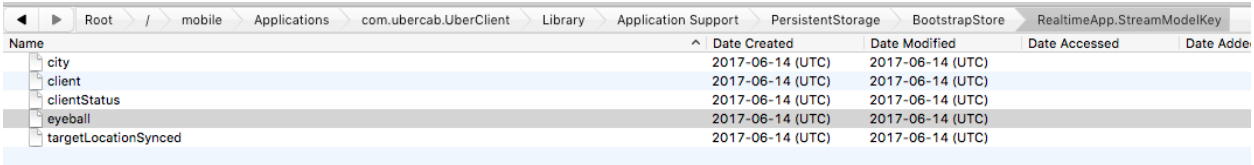


Figure 5. File in Uber application called "eyeball"

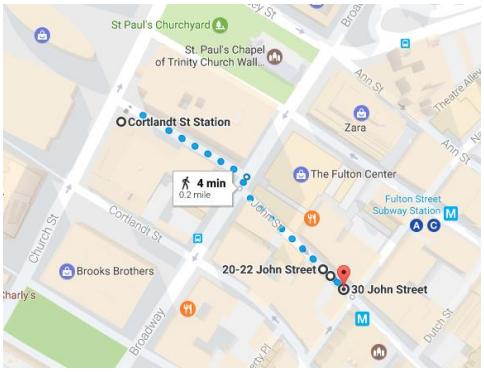


Figure 6. Vehicle path displayed in the Uber application

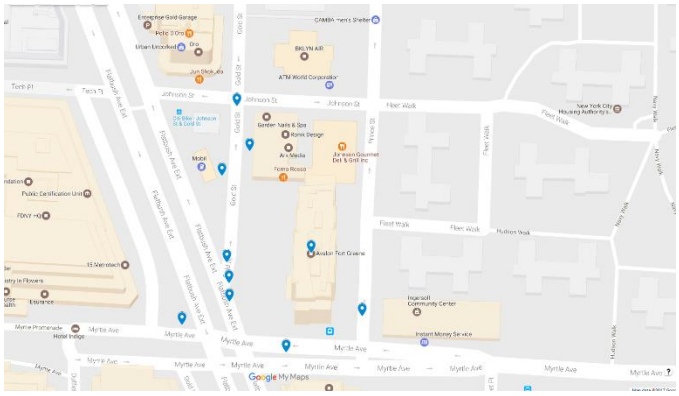


Figure 7. Map in Uber application tracking user in a Yellow Cab

latitude":40.69445117023311,"longitude":-73.98336568812392
types":
PICKUP"
id":"0c2aa2ef-c209-4edf-ad78-6fdfe9cc9a39","coordinate":
latitude":40.69387146666666,"longitude":-73.98332017882406
types":
PICKUP"
id":"43cd6b78-5f9d-43fc-a525-6350a5429855","coordinate":
latitude":40.6939573311466,"longitude":-73.9833223
types":
DROPOFF"
id":"95c3a984-fada-4cea-a02a-6270e5fc3eea","coordinate":
latitude":40.69363824659038,"longitude":-73.98297336909374
types":
DROPOFF"
id":"c0ec6651-2317-4b6a-93cc-7063b219b580","coordinate":
latitude":40.69456614993538,"longitude":-73.98319811377644
types":
DROPOFF"
id":"72837636-6f12-4ab7-b8ac-0548d1020169","coordinate":
latitude":40.69380613248446,"longitude":-73.9825112
types":
DROPOFF"
id":"d397aa0e-fa88-47ad-84f6-cd370cb8a504","coordinate":
latitude":40.69376409253024,"longitude":-73.98361030000001
types":
DROPOFF"
id":"97e0ed75-e849-466d-8ec2-b97f914c623d","coordinate":
latitude":40.69477418412908,"longitude":-73.98327546537712
types":
DROPOFF"
id":"7149ec60-cf05-4f59-8c45-d4de9576c2d5","coordinate":
latitude":40.69405269488668,"longitude":-73.98333687222757
types":
DROPOFF"
fullAddress":"343 Gold St, Brooklyn, New York 11201, US","addr
confidence":"HIGH"

Figure 8. "PICKUP" and "DROPOFF", with a Yellow Cab, recorded by the Uber app

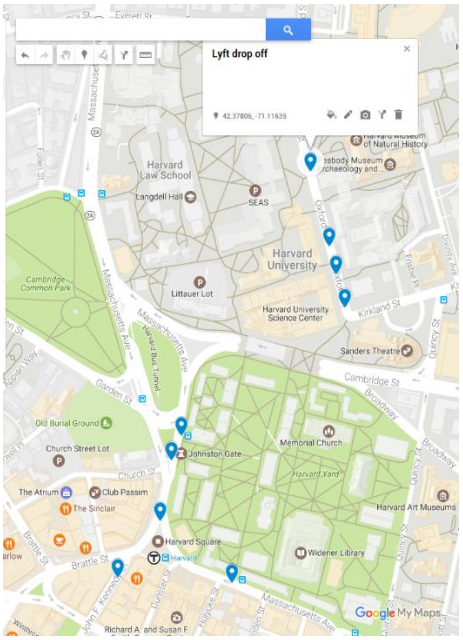


Figure 9. "PICKUP" and "DROPOFF", with the Lyft app, recorded by the Uber app

id": 050b4ecc-a724-443c-a07c-0a3f44c0e4b9", "coordinate":
latitude":42.37691095603864,"longitude":-71.11595316087985
types":
PICKUP"
id":"b63191d9-ba0b-4604-82e7-86b197601898","coordinate":
latitude":42.37641233312034,"longitude":-71.11577294999999
types":
DROPOFF"
id":"fd9d8dab-397f-4438-9f1e-e28eb41e70a2","coordinate":
latitude":42.37484584999999,"longitude":-71.11858413770518
types":
DROPOFF"
id":"a8fba781-34b1-4d59-b72f-527520179571","coordinate":
latitude":42.3731217492533,"longitude":-71.11967575576431
types":
DROPOFF"
id":"65f5b9fd-ee26-455a-b17c-e8cdba7e6478","coordinate":
latitude":42.37806711513722,"longitude":-71.11635314782401
types":
DROPOFF"
id":"c333291b-6e52-4aa2-81f9-c48bcc4e59da","coordinate":
latitude":42.37715454854897,"longitude":-71.11603882070287
types":
DROPOFF"
id":"ef65c525-6e49-4ec2-b8ad-d04443c79050","coordinate":
latitude":42.37681337220094,"longitude":-71.11591665623264
types":
DROPOFF"
id":"387ca06b-a122-4f56-b61a-b2454c086ebe","coordinate":
latitude":42.37454603113112,"longitude":-71.11875480934448
types":
DROPOFF"
id":"12640a66-a88f-4738-81c6-0ab17f70e9f5","coordinate":
latitude":42.374184,"longitude":-71.11883856666664
types":
DROPOFF"
id":"ca0a42b3-e92d-447c-a3b5-fbfea81dcc42","coordinate":
latitude":42.37381104686747,"longitude":-71.11894204606205
types":
DROPOFF"
id":"ce160ef6-7daa-4177-b70f-5e96d4ea697a","coordinate":
latitude":42.3776638023293,"longitude":-71.11622544012137
types":
DROPOFF"
id":"06d3824b-ae45-4cd2-a3cb-c7c7e3bf2527","coordinate":
latitude":42.37306188330274,"longitude":-71.11770618713565
types":
DROPOFF"
fullAddress":"Cambridge, MA 02138, United States","addressLine1":"Harvard University","provider":"google_places","addressLine2":"Cambridge, MA"
confidence":"HIGH"
Harvard UniversityCambridge, MA
y'/Y.
EjE0NTYxIEFtYm95IFJvYWQsIFN0YXRlbiBJc2xhbmQsIE5ZLCBVbml0ZWQgU3RhdGVz
analytics":
dataStream":"TEXT_SEARCH","dataSource":"UNKNOWN"

Figure 10. "PICKUP" and "DROPOFF", with the Lyft app, recorded by the Uber app

Facebook Fanatics: A Linguistic and Sentiment Analysis of the Most "Fanned" Facebook Pages

Alan R. Peslak
arp14@psu.edu
Penn State University
Dunmore PA 18512

Abstract

With nearly 2 billion users worldwide, Facebook is the most popular social media site in the world. Despite this popularity and ubiquity, it has been lightly studied in the literature. Our manuscript examines the most popular Facebook sites (pages) in the United States dealing with society and performs a comprehensive linguistics and sentiment analysis on these sites. Using Azure machine learning for sentiment and LIWC (Linguistic Inquiry and Word Count) for linguistics, our review finds significant similarities and differences in posts on Facebook pages that have the most fans (most popular). Implications and opportunities for further research are presented.

Keywords: Sentiment analysis, Facebook, Linguistic analysis, LIWC

1. INTRODUCTION

Facebook is the most popular social media site in the world. According to Zephoria (2017) in March of 2017, there are over 1.94 billion monthly active Facebook users. This is an 18 percent increase from the prior year. Every 60 seconds on Facebook: 510,000 comments are posted, 293,000 statuses are updated, and 136,000 photos are uploaded. One in five page views in the United States occurs on Facebook.

Facebook is perhaps the most popular form of communication in the world after verbal and telephone communications. It may be the most popular form of written communications in the world.

Because of its ubiquity and popularity it is a ripe area for research and analysis. Our manuscript analyzes one area of this massive communications vehicle, popular societal Facebook pages. Our analysis is to review posts on the most popular Facebook pages through linguistic and sentiment analysis. Our measure of popularity is based on Facebook "fans". According to Chan (2009) "In the same way that profiles on Facebook help you connect with friends, Facebook Pages allow you to interact with and stay up-to-date on your favorite public figures, organizations and businesses. When you become a fan of a Page, you are connecting with that organization or public figure and will begin seeing status updates, photos, videos and other posts from the Page. All of the posts from Pages will appear in your home page just as they would from your friends. You can get access to videos from your favorite band, chat live with your favorite celebrity, or even get a sneak peek of new products being launched by your favorite brand through Facebook Pages."

Linguistic and sentiment analysis is the review of written or verbal communications to determine specific characteristics of a communication. These characteristics can determine specific insights or meanings within a document, message, or speech that goes beyond the simple words in the communication. As a result, they can provide deeper understanding of the communicator's intent, bias, or personality thus framing the communication in a specific context and further clarify the communicator's full message. Our study reviews postings on the popular Facebook pages through linguistic and sentiment analysis.

2. LITERATURE REVIEW

Sentiment evaluation and linguistic analysis are commonplace techniques of studies in conversation analysis. The utilization of linguistic analysis and specially the use of LIWC (Linguistic and Word Count) software program for research functions has been substantial. Back, Kufner, and Egloff (2011) analyzed 11th of September communications the usage of LIWC. Cordova, Cunningham, Carlson, and Andrkowski (2001) used LIWC to research how individuals adjusted to having breast cancers. Robinson, Navea, and Ickes (2013) used LIWC analysis of college students written self-introductions to correctly calculate course performance. Bell, McCarthy, and McNamara (2012) used LIWC to analyze gender variations in linguistic styles. Sexton and Helmreich (2000) studied airline cockpit communications via LIWC to determine mistakes and overall performance. There are many other examples of the usage of LIWC inside the literature. The use of LIWC has been properly established and customary in peer-reviewed journals.

LIWC software (Pennebaker, Booth, Boyd, and Francis, 2015) is the most researched and popular linguistic analysis tool. "The way that the **L**inguistic **I**nquiry and **W**ord **C**ount program works is fairly simple. Basically, it reads a given text and counts the percentage of words that reflect different emotions, thinking styles, social concerns, and even parts of speech. Because LIWC was developed by researchers with interests in social, clinical, health, and cognitive psychology, the language categories were created to capture people's social and psychological states. …. The text analysis module then compares each word in the text against a user-defined dictionary. As described below, the dictionary identifies which words are associated with which psychologically-relevant categories." (Pennebaker Conglomerates, 2015).

Both Sentiment Analysis on Facebook posts and Linguistic analysis using LIWC have been used before in the literature. Kramer (2012) studied Facebook posts via LIWC and found that emotional status updates led to higher valence-consistent posts or posts that had more emotion. Getty et. Al (2011) studied deceased persons' Facebook profile posts via LIWC and found that Facebook served as a first stage grieving mechanism as well as maintaining a bond with the deceased. Farnadi et al. (2013) reviewed Facebook posts via LIWC to determine specific personality traits of individuals.

There has been limited study of sentiment analysis within Facebook posts. LIWC specifically does not measure sentiment. Troussas et al. (2013) examined sentiment of Facebook statuses and suggested a Naive Bayes classifier for language learning. Ortigosa et al. (2014) found studied Facebook posts to determine users' sentiment polarity with the goal of tailoring elearning systems based on students' sentiments.

One of the seminal studies in Sentiment Analysis is Sentiment Analysis and Opinion Mining by Bing Liu (2012). "Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, " Sentiment Analysis is the review of written or other forms of communication or qualitative data to determine a quantifiable and comparable measure of some form of feeling in the communication or data.

Pang and Lee (2008) deal with sentiment polarity and degree of positivity. A basic approach is to determine whether a particular communication is positive or negative. Eguchi and Lavrenko (2006) show this by retrieving selected sentiment text. "One of the first and still most used method of sentiment analysis is keyword analysis, "

3. METHODOLOGY

On March 16, 2017, a selection and review of the most popular Facebook pages was conducted. The website and company SocialBakers was used as the source of the most popular Facebook pages in the US according to number of Fans.

The purpose of the manuscript was to examine societal trends and not mere popular entertainment or games. We decided to focus on Facebook pages dedicated to Society. In order to examine potential differences due to types of Facebook pages, an overall society review based on societal categories was performed. The categories included and selected within SocialBakers were Politics, Science, CSR or Consumer Social Responsibility, Education, NGO or Non-Governmental Associations, and Professional Associations. In total Facebook posts in these six categories were reviewed and analyzed for similarities and differences.

In order to obtain posts from these popular pages, a Facebook mining tools known as Facepager was used. "Facepager was designed for fetching public available data from Facebook, Twitter and other JSON-based API. All data is stored in a local SQLite database and may be exported to csv." Appendix Figure 1 shows an

example output screen from Facepager. Facepager was invented by Till Keyling in 2011 and is actively developed and maintained by Jakob Jünger & Till Keyling. It is free of charge and open-sourced. (Keyling, Till; Jünger, Jakob, 2013)

The most popular pages in each categories and number of fans are listed in Table 1.

Politics	Barack Obama	54,588,709.00
	Donald J. Trump	21,528,119.00
Science	NASA	18,884,690.00
	NASA Sun Science	1,380,111.00
	NASA Universe Education	1,249,453.00
CSR	Johnson and Johnson Care Inspires Care	2,991,261.00
	My Black is Beautiful	2,621,868.00
Education	Harvard University	4,916,532.00
	Make Up First School of Makeup	3,937,150.00
NGO	Causes.com	8,638,052.00
	The Animal Rescue Site	7,858,016.00
Professional Associations	American Kennel Club	3,602,312.00
	USCCA	2,016,897.00

Table 1 Most Fanned FB Pages

The most recent 500 posts were retrieved for each Facebook page and analyzed (note that the Johnson and Johnson post only had 296 posts on the page and there was an unreadable post on American Kennel Club dropping their sample to 499).

Added to the spreadsheet was a sentiment variable calculated by the publicly available sentiment calculator from Microsoft Azure Machine Learning. This variable develops an overall measure of sentiment ranging from 0 (negative) to 1 (positive) with .5 being neutral. A specific multi-decimal rating is developed e.g. .54678 from this Microsoft Excel plug-in.

We also imported the posts into LIWC (Linguistic and Word Count). LIWC software results produce 93 unique measures from each of its linguistic analyses. These measures range from parts of speech to emotional categories to word counts. For the most part these are expressed by a percentage of total words mapping to the dictionary category of each measure. The exceptions are several relating to word counts as well as calculated emotional measures. Appendix Table 1 lists the LIWC variables used. There is also a definition of each or examples of words that meet the LIWC category. One-way ANOVA was performed to find differences among all 11 pages measured as well as between pages within a category (such as Donald Trump versus Barack Obama in politics). IBM SPSS 23.0 was used to develop the ANOVA and the results follow.

4. RESULTS

Page	Score	Category	Sig. Cat.
Donald J. Trump	65.07	Politics	0.293
Barack Obama	63.50	Politics	
NASA	66.48	Science	0.000
NASA Sun Science	60.97	Science	
NASA Universe Education	69.21	Science	
Johnson and Johnson Care Inspires Care	82.33	CSR	0.194
My Black is Beautiful	80.74	CSR	
Harvard University	65.03	Education	0.000
Make Up First School of Makeup	75.72	Education	
American Kennel Club	75.01	Prof. Assc.	0.001
USCCA	70.56	Prof. Assc.	
Total	69.59		0.000

Table 2 Sentiment as measured by Microsoft Azure

As noted the Facepager posts were analyzed via Microsoft Azure Learning sentiment analysis. The table two presents the results of the analysis. This and all scores are presented on a 1-100 scale with 0 being lowest and 100 highest. For sentiment, 1 would indicate very negative valence or sentiment, i.e. negative or bad feelings; 100 would represent very positive valence or sentiment, i.e. positive or good feelings. The tables show the sentiment score for each FB page. In addition, they show the societal category

they are attributed to and finally the significance category (Sig. Cat.) is the statistical significance of the difference between the two or more pages in that category. (This is the fourth column which is **bolded**). For example, the sentiment scores of Donald J. Trump and Barack Obama are .650746 and .634953 respectively. This is a small difference and is not statistically significant at $p < .05$. In fact p is actually **.293**, well above the .05 threshold. We can therefore conclude that there is no statistical difference between posts on Donald j. Trump's FB page and Barack Obama's FB page.

Page	Score	Category	Sig. Cat.
Donald J. Trump	75.47	Politics	.936
Barack Obama	75.32	Politics	
NASA	89.76	Science	.002
NASA Sun Science	90.04	Science	
NASA Universe Education	86.17	Science	
Johnson and Johnson Care Inspires Care	73.61	CSR	.696
My Black is Beautiful	74.48	CSR	
Harvard University	84.46	Education	.405
Make Up First School of Makeup	85.69	Education	
American Kennel Club	74.61	Prof. Assc.	0.001
USCCA	68.21	Prof. Assc.	
Total	80.04		**0.000**

Table 3 Analytic

The results of the Sentiment analysis overall present interesting results. First, the total sentiment across all selected pages and categories are all generally positive at 69 out of 100 but are significantly different at $p < .001$. The positive score is not surprising since these are pages that were "fanned" by FB users. The most positive scores were for CSR (Consumer Social Responsibility) and the two popular pages Johnson and Johnson Care Inspires Care and My Black is Beautiful were not significantly different with a p value of .194. Both scored above 80. Perhaps surprisingly the 65 and 63 sentiment scores of Trump and Obama were not significantly different. The diverse educational institutions of Harvard and Make Up First School of Makeup were significantly different with Harvard very much lower than Make Up. The fans of both the American Kennel Association and the United States Concealed Carry Association both have strong positive sentiments but AKA was significantly higher than USCCA. Finally, Science scores were somewhat lower than expected and significantly different.

The rest of the analyses all use the results of LIWC. According to (Pennebaker, Booth, Boyd, and Francis, 2015) "Analytical thinking ------ a high number reflects formal, logical, and hierarchical thinking; lower numbers reflect more informal, personal, here --- and --- now, and narrative thinking.".

Page	Score	Category	Sig. Cat.
Donald J. Trump	68.31	Politics	.875
Barack Obama	68.03	Politics	
NASA	67.08	Science	0.00
NASA Sun Science	56.08	Science	
NASA Universe Education	68.15	Science	
Johnson and Johnson Care Inspires Care	85.64	CSR	.532
My Black is Beautiful	86.73	CSR	
Harvard University	66.64	Educ	.164
Make Up First School of Makeup	64.58	Educatio n	
American Kennel Club	76.96	Prof. Assc.	.029
USCCA	80.32	Prof. Assc.	
Total	71.14		0.000

Table 4 Clout

For the most part, all society FB pages showed high analytic content (overall averaging 73.82). Not surprisingly, the most analytic were the NASA Science posts. The least analytic was the US Concealed Carry Association at 68.21. The least analytic overall category was Professional Associations, mostly due to the USCCA. The results indicate that for society issues posts are generally formal and logical not emotional. This is perhaps a surprising result since FB is often seen as a casual and informal means of communication, as in Lofters, A. K., Slater, M. B., Nicholas Angl, E., & Leung, F.-H. (2016). This

ranking suggests that FB may provide a higher level of discourse than previously proposed.

The next measure analyzed was Clout.

"Clout ------ a high number suggests that the author is speaking from the perspective of high expertise and is confident; low Clout numbers suggest a more tentative, humble, even anxious style." (Pennebaker, Booth, Boyd, and Francis, 2015).

In general, all society FB pages have a high level of clout or level of confidence. The highest are the Consumer Social Responsibility pages. This suggests they are trusted and/or highly knowledgeable. Surprisingly, the lowest level of confidence was in Education and Science categories. This perhaps suggests the more inquisitive and exploratory nature related to these fields. It is interesting to note that posts on both Trump and Obama Politics pages are about average for these categories and are not significantly different in clout. The USCCA has the highest confidence after the CSR pages.

Page	Score	Category	Sig. Cat.
Donald J. Trump	37.45	Politics	.001
Barack Obama	30.11	Politics	
NASA	47.99	Science	.015
NASA Sun Science	50.62	Science	
NASA Universe Education	44.96	Science	
Johnson and Johnson Care Inspires Care	27.25	CSR	.243
My Black is Beautiful	29.96	CSR	
Harvard University	28.94	Education	0.000
Make Up First School of Makeup	15.46	Education	
American Kennel Club	21.60	Prof. Assc.	0.000
USCCA	28.77	Prof. Assc.	
Total	33.23		0.000

Table 5 Authentic

The authenticity (Authentic) measure averages only 30.44. "Higher numbers are associated with a more honest, personal, and disclosing text; lower numbers suggest a more guarded, distanced form of discourse." (Pennebaker, Booth, Boyd, and Francis, 2015). Our results suggest a less personal form of disclosure and more distanced.

LIWC Analytic Measures for Selected Media (Pennebaker, J., Boyd, R., Jordan, K., and Blackburn, K. (2015))

Natural speech has been analyzed to be 61 on the 0-100 scale. (See Appendix Table 2 for LIWC ratings for common forms of communication). The Societal FB posts average only 30. This proposes that FB posts are much more guarded than natural speech. This may be due to the broad public nature of FB. People may be much more leery of posting comments on FB than natural speaking. This has broad implications for analyzing the content of FB posts. These posts may not reflect all true feelings. Due to a less controversial area, the Science posts are much less guarded.

Page	Score	Category	Sig. Cat.
Donald J. Trump	58.57	Politics	.024
Barack Obama	52.87	Politics	
NASA	43.75	Science	0.000
NASA Sun Science	52.61	Science	
NASA Universe Education	55.15	Science	
Johnson and Johnson Care Inspires Care	84.80	CSR	.020
My Black is Beautiful	79.48	CSR	
Harvard University	51.68	Education	.228
Make Up First School of Makeup	48.93	Education	
American Kennel Club	71.58	Prof. Assc.	0.000
USCCA	54.46	Prof. Assc.	
Total	58.47		0.000

Table 6 Tone

Tone differs somewhat from sentiment. "Emotional tone ------ a high number is associated with a more positive, upbeat style; a low number reveals greater anxiety, sadness, or hostility. A number around 50 suggests either a lack of

emotionality or different levels of ambivalence." The emphasis here is on style, upbeat or sad. Overall the average was slightly more than ambivalent. Trump posts were significantly more upbeat than Obama. The CSR pages were very upbeat as was the American Kennel Club. NASA as a whole was a bit more downbeat than the NASA specific sites. The rest hovered around the non-emotional average. None of these results seem surprising and are consistent with their messages and followers.

Page	Score	Category	Sig. Cat.
Donald J. Trump	8.29	Politics	.000
Barack Obama	6.54	Politics	
NASA	5.36	Science	0.000
NASA Sun Science	5.02	Science	
NASA Universe Education	8.87	Science	
Johnson and Johnson Care Inspires Care	11.37	CSR	.085
My Black is Beautiful	12.40	CSR	
Harvard University	5.69	Education	0.000
Make Up First School of Makeup	4.16	Education	
American Kennel Club	8.78	Prof. Assc.	0.000
USCCA	11.64	Prof. Assc.	
Total	7.88		0.000

Table 7 Pronoun

A high percentage of pronouns reveals a more personal and informal style. Natural speech rates at 15% usage of pronouns. All our Facebook Societal posts are less than this, averaging only 7.97. The most personal were in CSR posts at between 11 and 12. The lowest were Science and Education at 5. In general, it can be said that Societal FB page posts reveal little of personal style. This is consistent with the authenticity measure which showed more guarded communications.

Page	Score	Category	Sig. Cat.
Trump	5.96	Politics	0.011
Barack Obama	4.79	Politics	
NASA	1.50	Science	0.000
NASA Sun Science	2.49	Science	
NASA Universe Education	2.54	Science	
Johnson and Johnson Care Inspires Care	9.83	CSR	0.000
My Black is Beautiful	7.41	CSR	
Harvard University	3.59	Education	0.000
Make Up First School of Makeup	2.34	Education	
American Kennel Club	7.86	Prof. Assc.	0.000
USCCA	4.19	Prof. Assc.	
Total	4.58		0.000

Table 8 Positive emotions

Page	Score	Category	Sig. Cat.
Trump	1.43	Politics	0.798
Barack Obama	1.48	Politics	
NASA	0.38	Science	0.535
NASA Sun Science	0.32	Science	
NASA Universe Education	0.34	Science	
Johnson and Johnson Care Inspires Care	0.13	CSR	0.045
My Black is Beautiful	0.32	CSR	
Harvard University	0.93	Education	0.000
Make Up First School of Makeup	0.26	Education	
American Kennel Club	0.66	Prof. Assc.	0.000
USCCA	1.58	Prof. Assc.	
Total	.73		0.000

Table 9 Negative emotions

Natural speech tends to average about 5% positive emotions and 1% negative emotions. The results of the FB post analysis suggests that FB posts for the most popular "Fanned" FB pages mirrors this general level and ratio. The actual results are 4.55 positive emotions and .81 negative emotions. Thus, generally, emotionally we communicate similarly on FB fan pages and in natural speech. If confirmed through further studies this is an important finding with regard to online versus in-person communications.

Also there are very interesting results with regard to individual pages and categories. In Politics, Donald J. Trump fans have statistically significant higher positive emotions than Barack Obama, but show no difference in negative emotions. Science also has significant differences in positive emotions but not in negative emotions. All the Science pages show little positive emotions though, with NASA almost 0 at 1.5%. CSR pages have much higher positive emotions and much lower negative emotions. There are significant differences in Johnson and My Black however. Johnson has higher positive and lower negative.

Page	Category	Ratio	Average
Donald J. Trump	**Politics**	4.16	3.702
Barack Obama	Politics	3.23	
NASA	**Science**	3.94	6.39
NASA Sun Science	Science	7.78	
NASA Universe Education	Science	7.47	
Johnson and Johnson Care Inspires Care	**CSR**	75.61	49.38
My Black is Beautiful	CSR	23.15	
Harvard University	**Education**	3.86	6.43
Make Up First School of Makeup	Education	9	
American Kennel Club	**Prof. Assc.**	11.90	7.28
USCCA	Prof. Assc.	2.65	
Total		5.61	

Table 10 Positive to negative emotions ratio

Education has significantly higher positive emotions for Harvard but also higher negative than Make Up. Finally, the American Kennel Club has significantly higher positive emotions and lower negative emotions than US Concealed Carry.

5. FURTHER DISCUSSION

Overall, the sentiment and linguistic analyses of popular FB pages yields some very stimulating results. The sentiment measure for Facebook pages that individuals have fanned are generally positive. Scores ranged from 65 to 82 with Politics on the lower end and Consumer Social Responsibility on the higher end. This sentiment analysis is somewhat confirmed by the positive and negative emotion ratings that LIWC has developed. The highest positive emotions were for CSR but the lowest positive emotions are for the Science category. Further analysis of the ratio of positive to negative emotions however better supports overall sentiment.

Politics actually had the lowest ratio of positive to negative emotions of any category at 3.70. Science was actually above total at 6.39. CSR had an incredibly high positive to negative ratio of 49.4.

Comparing this ratio to common other forms of communication also yields interesting results.

	Positive emotions	Negative emotions	Ratio
Blogs	3.66	2.06	1.78
Expressive writing	2.57	2.12	1.21
Novels	2.67	2.08	1.28
Natural Speech	5.31	1.19	4.46
NY Times	2.32	1.45	1.60
Twitter	5.48	2.14	2.56

Table 11 Other Communications Positive to negative emotions ratio

For all Facebook posts we analyzed the ratio of positive to negative emotions compared to Blogs, expressive writing, novels, natural speech, NY Times, and Twitter. The FB posts are significantly more positive in emotion than all these forms of communications. It is posited then that these FB posts do not accurately reflect how we communicate in everyday life or other forms of communication. Rather they are artificially positive in their message and content and provide

support and good feelings not consistent with other forms of discourse.

All the FB pages studied scored high in analytic measures ranging from 90 for NASA to 68 for USCCA. These measures were over all other forms of communication in Appendix Table 2 except the New York Times. Posts were generally formal and logical. In general, confidence levels of the reviewed FB posts were confident and strong. Scores ranged from a high of 80 for USCCA to 56 for NASA Sun Science. By comparison, regular speech only has a 52 Clout measure and the New York Times is only 68. Coincidentally, the Clout scores for Donald J. Trump and Barack Obama were also 68. Level of authenticity for all FB posts was low; much lower than natural speech, novels, or blogs (61, 75, 60 respectively). The FB posts ranged from a low of 15 for Make Up School to a high of 50 for NASA Sun Science. All others were below 50. This suggests a level of guardedness in what is posted online. Many are near the guardedness of the New York Times which is at 25. Overall Tone of the posts for the most part was somewhat neutral, though the CSR posts were very upbeat at 80 and above. Interestingly, natural speech occurs at about this 80 level, much higher than other forms of communication such as blogs (55), the New York Times (44) and novels (37) as well as most of our studied FB posts (average 58). The personal aspect of our FB posts as measured by the use of pronouns was for the most part much less than natural speech (15) or the New York Times (21). Generally it can be said that FB posts are less personal than most other forms of communication.

6. CONCLUSION

It should be noted that there are limitations to the study. First, only one day was selected but many posts were prior to that day. Further duplication studies should be performed over time. Next, only 500 posts were used for each analysis. Greater numbers may yield different results. Finally only top "fanned" pages were used. Different results may be obtained by using lesser popular pages.
Despite these limitations, this study has demonstrated a series of important results. This study of sentiment analysis extends the work of many applied IS research including highly cited works from Computers in Human Behavior, Communications of the ACM, SIGCHI conferences and Expert Systems with applications, *Foundations and Trends® in Information Retrieval.* First it defines, presents and demonstrates an example and interpretation of linguistic analysis and sentiment analysis using one of the most researched and developed tools, LIWC. Researchers and practitioners can use this manuscript as a source and guide for developing their own linguistic analysis of any communication. Second, the study illustrates the results of Facebook posts metrics as they compare to other forms of computer-mediated communications. Researchers and practitioners can reliably use this comparison for other forms of computer-mediated communications. Finally, the study analyzes Facebook posts via linguistic and sentiment of the most popular FB Society sites and categories. The results show significant differences in all areas of sentiment and linguistic analyses. There are also significant differences within categories. Researchers can use these findings to compare and contrast Facebook posts to their linguistic characteristics. Societal social network Facebook page hosts can use these findings to improve their overall sentiment and linguistic metrics if they choose.

7. REFERENCES

Back, M. D., Küfner, A. C., & Egloff, B. (2011). Automatic or the people? Anger on September 11, 2001, and lessons learned for the analysis of large digital data sets. *Psychological Science*, *22*(6), 837-838.

Bell, C. M., McCarthy, P. M., & McNamara, D. S. (2012). Using LIWC and Coh-Metrix to investigate gender differences in linguistic styles. *Applied Natural Language Processing: Identification, Investigation, and Resolution, Information Science Reference, Hershey, PA*, 545-556.

Chan, K. (2009) Facebook Tips: How do I find and "fan" a Page? https://www.facebook.com/notes/facebook/facebook-tips-how-do-i-find-and-fan-a-page/185405397130/

Cordova, M. J., Cunningham, L. L., Carlson, C. R., & Andrykowski, M. A. (2001). Social constraints, cognitive processing, and adjustment to breast cancer. *Journal of consulting and clinical psychology*, *69*(4), 706.

Eguchi and V. Lavrenko, "Sentiment retrieval using generative models,"in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 345–354, 2006

Farnadi, G., Zoghbi, S., Moens, M. F., & De Cock, M. (2013, January). Recognising personality

traits using facebook status updates. In *Proceedings of the workshop on computational personality recognition (WCPR13) at the 7th international AAAI conference on weblogs and social media (ICWSM13)*. AAAI.

Getty, E., Cobb, J., Gabeler, M., Nelson, C., Weng, E., & Hancock, J. (2011, May). I said your name in an empty room: grieving and continuing bonds on facebook. In *Proceedings of the SIGCHI Conference on human factors in computing systems* (pp. 997-1000). ACM.

Keyling, Till; Jünger, Jakob (2013). Facepager (Version, f.e. 3.3). An application for generic data retrieval through APIs. Source code available from https://github.com/strohne/Facepager.

Kramer, A. D. (2012, May). The spread of emotion via Facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 767-770). ACM.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Lofters, A. K., Slater, M. B., Nicholas Angl, E., & Leung, F.-H. (2016). Facebook as a tool for communication, collaboration, and informal knowledge exchange among members of a multisite family health team. *Journal of Multidisciplinary Healthcare*, *9*, 29–34. http://doi.org/10.2147/JMDH.S94676

Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*, *31*, 527-541.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, *2*(1-2), 1-135.

Pennebaker Conglomerates (2015) *LIWC How It Works*. http://liwc.wpengine.com/how-it-works/

Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015*). Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX : Pennebaker Conglomerates (www.LIWC. net).,

Pennebaker, J., Boyd, R., Jordan, K., and Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX. University of Texas at Austin.

Robinson, R. L., Navea, R., & Ickes, W. (2013). Predicting final course performance from students' written self-introductions: A LIWC analysis. *Journal of Language and Social Psychology*, 0261927X13476869.

Sexton, J. B., & Helmreich, R. L. (2000). Analyzing cockpit communications: the links between language, performance, error, and workload. *Journal of Human Performance in Extreme Environments*, *5*(1), 6.

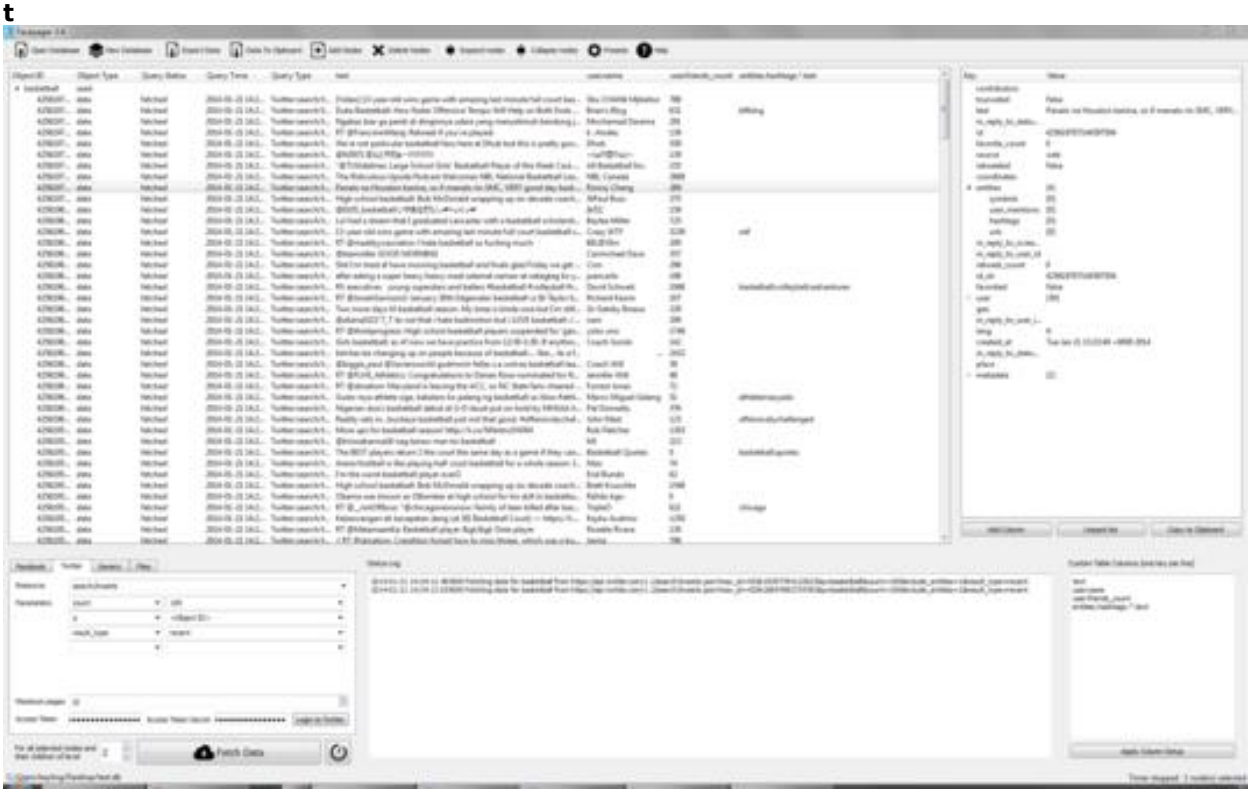
SocialBakers. (2017) Top Facebook pageshttps://www.socialbakers.com/statistics/facebook/pages/total/united-states/society/science/

Troussas, C., Virvou, M., Espinosa, K. J., Llaguno, K., & Caro, J. (2013, July). Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning. In *Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference on* (pp. 1-6). IEEE.

Zephoria (2017) https://zephoria.com/top-15-valuable-facebook-statistics/

Appendices

Appendix Figure 1 Sample Facepager outpu t



Appendix Table 1 Dependent and Independent variable table

Variable	Meaning or examples
SentN	Overall sentiment of tweet (0 negative to 1 positive)
Analytic	reflects logical thinking versus narrative
Clout	Confident (high score) versus Tentative (low score)
Authentic	Honest versus Guarded
Tone	Upbeat versus sad
Dic	Number of words in LIWC dictionary (suggests ease of read)
Pronoun	I, them, itself
Posemo	love, nice, sweet
Negemo	hurt, ugly, nasty

Appendix Table 2. LIWC Analytic Measures for Selected Media (Pennebaker, J., Boyd, R., Jordan, K., and Blackburn, K. (2015))

	Blogs	Expressive writing	Novels	Natural Speech	NY Times	Twitter	Explanation
Analytic	49.89	44.88	70.33	18.43	92.57	61.94	Logical versus informal
Clout	47.87	37.02	75.37	52.27	68.17	63.02	Confident versus humble
Authentic	60.93	76.01	21.56	61.32	24.84	50.39	Honest versus guarded
Tone	54.5	38.6	37.06	79.29	43.51	72.24	Upbeat versus hostile
Dictionary	85.79	91.93	84.52	91.6	74.62	82.6	Nontechnical, Number of words in LIWC dictionary (suggests ease of read)
Pronoun	16.2	16.2	18.03	15.15	20.92	7.41	Personal and informal
Positive emotions	3.66	2.57	2.67	5.31	2.32	5.48	Happy
Negative emotions	2.06	2.12	2.08	1.19	1.45	2.14	Sad or angry

Minimalism in Data Visualization: Perceptions of Beauty, Clarity, Effectiveness, and Simplicity

Stephen Hill
hills@uncw.edu

Barry Wray
wrayb@uncw.edu

Christopher Sibona
sibonac@uncw.edu

Business Analytics, Information Systems, and Supply Chain
University of North Carolina Wilmington
Wilmington, NC, 28403-5611, USA

Abstract

Minimalism in data visualization has been espoused by experts such as Edward Tufte for many years. In this work, minimalism in basic charts, as represented by the data-ink ratio in those charts is examined. A survey was developed and respondents were asked to provide their perceptions of a series of barplots and scatterplots on the dimensions of beauty, clarity, effectiveness, and simplicity. Differing data-ink ratios were presented in the charts, with high data-ink ratio charts representing minimalist design. Analysis of the survey respondents' perceptions suggested that visualizations with lower data-ink ratios were better on each of the dimensions. This finding is in contrast to the philosophy espoused by Tufte and is in line the findings of other previous work. Conclusions, discussion, and opportunities for future work are provided.

Keywords: Visualization, Minimalism, Perception, Analytics

1. INTRODUCTION

Whether it be in the news media, academic publications, or student reports, visualizations of data are commonplace. It is then of critical importance that the visualizations that are part of this increased prevalence be subject to scrutiny with regard to design and effectiveness. The focus of this work is an analysis of the perceptions surrounding the application of minimalist design techniques to data visualizations.

Retired Yale professor Edward Tufte has gained a widespread reputation and following for his data visualization expertise. One of his key visualization philosophies is the application of minimalism. This philosophy is described in detail in Tufte's well-known text, *The Visual Display of Quantitative Information* (2001). Tufte's minimalist philosophy is rooted in the concept of the data-ink ratio. This ratio is defined as the ratio of "ink" that is used in a graphic to display data to the total ink used to print the graphic.

Figure 1 shows an example of low and high data-ink ratio visualizations (Haims, 2012). As the data-ink ratio in a visualization is increased, the visualization becomes more minimalist in style. A high data-ink ratio in a visualization is considered, by Tufte, to be superior. Tufte goes further and refers to non-data ink or redundant ink as "chartjunk" (Tufte, 2001). Figure 2 shows a

graphic from Holmes (1984) that has been cited as a classic example of chartjunk (Few, 2011).

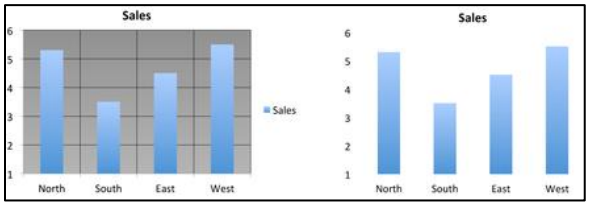


Figure 1: Data-Ink Ratio Examples (Low Ratio at Left, High Ratio at Right)

Figure 2: Chart Junk Example

The core question associated with this research is then: Is Tufte's minimalist approach actually superior? The concept of "superior" is evaluated in this research in the context of a visualization's beauty, clarity, effectiveness, and simplicity. Such an approach to visualization evaluation is not entirely new, but this work contributes to the literature by building upon the work of Inbar et al. (2007). In particular, this work adds consideration of scatterplots and an effectiveness evaluation dimension.

This paper is organized as follows. A brief overview of the literature related to data visualization with a specific focus on the use of Tufte minimalism is provided. The methodology used in this work is then described. Survey results are presented and analyzed. The article closes with a discussion of the implications of the results and identifies opportunities for future work.

2. LITERATURE REVIEW

Data visualization and the creation of effective visualizations have been the focus of considerable work in the literature. For example, Heer et al. (2010) illustrate the wide variety of potential, modern visualizations available. Other popular, modern resources for data visualization include: Knaflic (2015), Few (2012), and Evergreen (2016).

The concept of minimalism in data visualizations has been discussed in the academic literature. The work most closely related to this work is that of Inbar et al. (2007). In that article, the researchers examined Tufte's philosophy of minimalism via the data-ink ratio. The 87 survey respondents (all undergraduate students in Israel) were presented four bar charts. The bar charts featured data-ink ratios ranging, subjectively, from low to high.

The respondents were asked to evaluate each chart on the basis of the dimensions of beauty, clarity, and simplicity. Each chart's effectiveness was not considered. The results suggested that respondents preferred charts with lower data-ink ratios and considered such charts to be more beautiful, clear, and simple. The researchers offer a limited view on the potential causes driving their results.

In their work, Blasio and Bisantz (2002) suggest that a high data-ink ratio may be effective in reducing the time needed to recognize an outlier event in a visualization. They provide experimental results to support their claim. Sorensen (1993) found that the presence of a background image (lower data-ink ratio) on a chart was detrimental to perceptions of chart quality. Hullman, Adar, and Shah (2011) suggest that the use of minimalist concepts are recommended when preparing graphics that may be viewed by those with accessibility/disability concerns.

However, Bateman et al. (2010) suggested that the use of "chart junk" (lower data-ink ratio) in the form of overt embellishment of charts resulted in better recall of the data and did not adversely impact interpretation of the data. Their work was driven, primarily, by subjective surveys with open-ended responses. Similar results were echoed in work by Li and Moacdieh (2014) and McGurgan (2015). Borkin et al. (2013) found that embellished graphs resulted in improved memorability over graphs that were considered to be plain or standard. Norman (2008) goes further than the others in plainly stating that "simplicity is not the answer". However, he provides no empirical evidence to support his assertion.

In summary, the academic literature on minimalism and data visualization, provides

conflicting empirical evidence related to minimalism's ability to improve a visualization. This debate has carried over outside of academia as evidenced by blog postings such as that of Kosara (2010) that discusses the work of Bateman et al. (2010). However, most of the literature seems to suggest that minimalism in the form of a high data-ink ratio or lack of embellishment does not improve the quality of a visualization. This is counter to Tufte's notion that such visualizations are superior.

3. METHODOLOGY

This section describes the development of the visualizations and survey used in this work.

Visualization Development

In contrast to Inbar et al's. (2007) consideration of only barplots, two types of charts were created for this work: barplots and scatterplots. A total of eight visualizations were developed, four for each of the two chart types. Plot A for both types of charts featured the lowest data-ink ratio. Moving from Plot A to Plot D corresponded to an increase in data-ink ratio (increased minimalism). Figures 3 and 4 in the Appendix show the four barplots and scatterplots, respectively. Table 1 shows the chart elements that were present in each of the plots. Note that none of the plots contained embellishments in the style of Holmes (1984). Rather the plots contained typical chart elements that might be found in default or near-default plots generated in Excel.

Survey Development

A ten-item survey was created to evaluate perceptions of the two sets of plots. The survey was composed of two blocks of four questions per block (corresponding to each set of plots), plus two questions concerning basic survey respondent demographics. Users were presented both survey blocks, but in random order. Within the blocks, four plots of the same type were shown to the user, with the order of plots within the blocks also randomized. The chart types (bar plot and scatterplot) were selected for their perceived popularity, particularly among Excel users.

The respondents were asked, on a seven-point Likert scale (Strongly Disagree to Strongly Agree with Neither Agree nor Disagree as a neutral option), to evaluate each of the eight plots on the basis of beauty, clarity, effectiveness, and simplicity. These terms were not defined for the respondents, so the respondents were able to develop their own subjective definitions or notions for these terms. Note that not providing definitions for these terms is consistent with Inbar et al's. (2007) approach.

	Plot			
Element	**A**	**B**	**C**	**D**
Chart Title	X			
Chart Shading	X			
x Axis Label	X	X	X	
y Axis Label	X	X	X	
Horizontal Gridlines	X	X		
Vertical Gridlines	X			
Data Fill	X	X	X	

Table 1: Elements of Charts Presented in Survey

Analysis

Data analysis of the survey results was conducted using the R statistical programming software (R Core Team, 2015) in the RStudio development environment (RStudio Team, 2015). The *reshape2* (Wickham, 2016b) and *plyr* (Wickham, 2016a) packages were used for data preparation. The R package *likert* was used to develop visualizations of the survey responses (Bryer, Speerschneider, & Bryer, 2016).

4. RESULTS

The survey for this work was distributed to 116 students enrolled in at least one of three courses (Introduction to Operations Management, Introduction to Statistics, and Professional Writing) taught during an academic term in 2017 at a large, regional university in the southeastern United States. Of these students, 70 (60.3%) submitted a survey response. Students were incentivized to complete the survey by being offered a small amount of bonus credit in their respective course.

The median age of the respondents was 20 years with a maximum age of 56 years and a minimum age of 18 years. Of the respondents, 38 were indicated that they were female and 32 indicated that they were male. Eleven of the students were enrolled in the Professional Writing course, 23 in Introduction to Operations Management, and 36 in Introduction to Statistics.

Figures 5 and 6 in the Appendix show the distributions of student survey responses for each of the plots (A, B, C, and D) for the barplots and

scatterplots, respectively. Responses are grouped by the Beauty, Clarity, Effectiveness, and Simplicity dimensions. These figures suggest that the survey respondents largely viewed the minimalist plots (Plots D) as the least beautiful, clear, effective, and simple. The respondents found the plots (Plots A), with the lowest data-ink ratio of the four plots, to be highest rated on each of the four dimensions.

For the barplots, plots A and B are perceived to be roughly equivalent on all four dimensions. Stripping both horizontal and vertical gridlines to create plot C results in a significant decrease (p value less than 0.05) as indicated by Fisher's Exact Test (Fisher, 1992), on all four dimensions. For the scatterplots, there are greater differences in perceptions between the four plots for the Beauty, Clarity, and Effectiveness dimensions. Each difference was significant (p value less than 0.05). Plots A, B, and C are perceived similarly on the Simplicity dimension.

When comparing the plots in aggregate (grouping responses associated with Plots A, B, C, and D together), survey respondents perceived that barplots were better than the scatterplots on each of the four dimensions. This difference was found to be significant (p values less 0.05) and is illustrated in Figure 7 in the Appendix.

Although the researchers did not preconceive that there would be differences in perceptions by gender, differences were found. Across all plots, female students perceived lower levels of beauty and effectiveness than their male counterparts. The difference was significant for the simplicity and effectiveness dimensions. Females tended to perceive each of the scatterplots as featuring greater simplicity than the males. Figures 8 and 9 in the Appendix illustrate these differences in perception. Gender differences in perception of plot beauty, clarity, effectiveness, and simplicity may be an interesting extension of this work.

5. CONCLUSIONS AND DISCUSSION

In this work, perceptions of minimalist approaches to data visualization were examined via a survey. Respondents were presented with two sets of four plots and were asked to rate each plot on perceived Beauty, Clarity, Effectiveness, and Simplicity. The plots presented similar data, but were varied in their data-ink ratio.

The results of the survey response analysis suggest that plots with lower data-ink ratios (less minimalist) are perceived as superior on each of the four dimensions. This held true whether the plots presented were barplots or scatterplots. These results are in largely in agreement with previous work by Inbar et al. (2007) on barplots across the dimensions of Beauty, Clarity, and Simplicity. The results are contrary to the philosophy of highly-regarded visualization expert Edward Tufte.

Despite these findings, questions remain. For example, why did the survey respondents prefer the plots with lower data-ink ratios? The authors of this work have formed a post-analysis hypothesis This hypothesis suggests that the respondents are "used" to seeing default plots as produced by Excel. These plots feature a moderately low data-ink ratio. For example, the default barplot (using the same data as in this work) from Excel is shown in Figure 10. Such a plot would be familiar to almost any user of Excel.

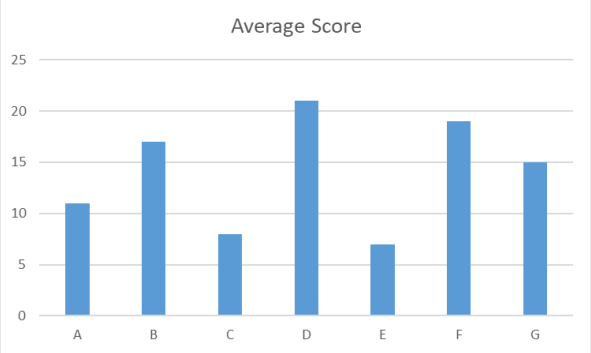


Figure 10: Default Excel Barplot

A lack of familiarity with minimalist visualizations with high data-ink ratios may drive respondents to view these visualizations less favorably. This hypothesis could provide an opportunity for future work. Also, survey respondents were not provided definitions for the terms "beauty", "clarity", "effectiveness", and "simplicity". Future work could examine the potential impact of providing survey respondents with formal definitions of these terms. Would reducing the subjectivity of these terms affect respondent's choices?

Additionally, expanding the survey response base to include additional chart types as well as respondents of varying age and educational backgrounds could lead to other findings. Furthermore, exploration of gender differences in perception may be of interest to future researchers.

6. REFERENCES

Bateman, S., Mandryk, R. L., Gutwin, C., Genest, A., McDine, D., & Brooks, C. (2010). Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2573–2582). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1753716

Blasio, A. J., & Bisantz, A. M. (2002). A comparison of the effects of data–ink ratio on performance with dynamic displays in a monitoring task. *International Journal of Industrial Ergonomics*, *30*(2), 89–101. https://doi.org/10.1016/S0169-8141(02)00074-4

Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What Makes a Visualization Memorable? *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2306–2315. https://doi.org/10.1109/TVCG.2013.234

Bryer, J., Speerschneider, K., & Bryer, M. J. (2016). Package "likert." Retrieved from http://cran.pau.edu.tr/web/packages/likert/likert.pdf

Evergreen, S. D. H. (2016). *Effective Data Visualization: The Right Chart for the Right Data*. SAGE Publications.

Few, S. (2011). The Chartjunk Debate. *Visual Business Intelligence Newsletter, No. June*, 1–11.

Few, S. (2012). *Show Me the Numbers: Designing Tables and Graphs to Enlighten* (2nd ed.). USA: Analytics Press.

Fisher, R. A. (1992). Statistical Methods for Research Workers. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics* (pp. 66–70). Springer New York. https://doi.org/10.1007/978-1-4612-4380-9_6

Haims, N. (2012, March 15). Data-Ink and The Dangers of Chart Junk in Information Design. Retrieved June 14, 2017, from http://nolanhaimscreative.com/2012315data-ink-and-the-dangers-of-chart-junk-in-information-design-html/

Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Commun. Acm*, *53*(6), 59–67.

Holmes, N. (1984). *Designer's Guide to Creating Charts and Diagrams*. New York, NY, USA: Watson-Guptill Publications.

Hullman, J., Adar, E., & Shah, P. (2011). Benefitting InfoVis with Visual Difficulties. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2213–2222. https://doi.org/10.1109/TVCG.2011.175

Inbar, O., Tractinsky, N., & Meyer, J. (2007). Minimalism in Information Visualization: Attitudes Towards Maximizing the Data-ink Ratio. In *Proceedings of the 14th European Conference on Cognitive Ergonomics: Invent! Explore!* (pp. 185–188). New York, NY, USA: ACM. https://doi.org/10.1145/1362550.1362587

Knaflic, C. N. (2015). *Storytelling with Data: A Data Visualization Guide for Business Professionals*. John Wiley & Sons.

Kosara, R. (2010, April 22). Chart Junk Considered Useful After All. Retrieved June 15, 2017, from https://eagereyes.org/criticism/chart-junk-considered-useful-after-all

Li, H., & Moacdieh, N. (2014). Is "chart junk" useful? An extended examination of visual embellishment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 1516–1520. https://doi.org/10.1177/1541931214581316

McGurgan, K. (2015). Data-ink Ratio and Task Complexity in Graph Comprehension. Retrieved from http://scholarworks.rit.edu/theses/8668/

Norman, D. A. (2008). The Way I See It: Simplicity is not the answer. *Interactions*, *15*(5), 45. https://doi.org/10.1145/1390085.1390094

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

RStudio Team. (2015). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc. Retrieved from http://www.rstudio.com/

Sorensen, D. L. (1993). *The Psychological Processes of Embellished Graph Reading*. University of Idaho, Moscow, ID.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Wickham, H. (2016a). plyr: Tools for Splitting, Applying and Combining Data (Version 1.8.4). Retrieved from https://cran.r-project.org/web/packages/plyr/index.html

Wickham, H. (2016b). reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package (Version 1.4.2). Retrieved from https://cran.r-project.org/web/packages/reshape2/index.html

Editor's Note:

This paper was selected for inclusion in the journal as a CONISAR 2017 Meritorious Paper. The acceptance rate is typically 15% for this category of paper based on blind reviews from six or more peers including three or more former best papers authors who did not submit a paper in 2017.

Appendix

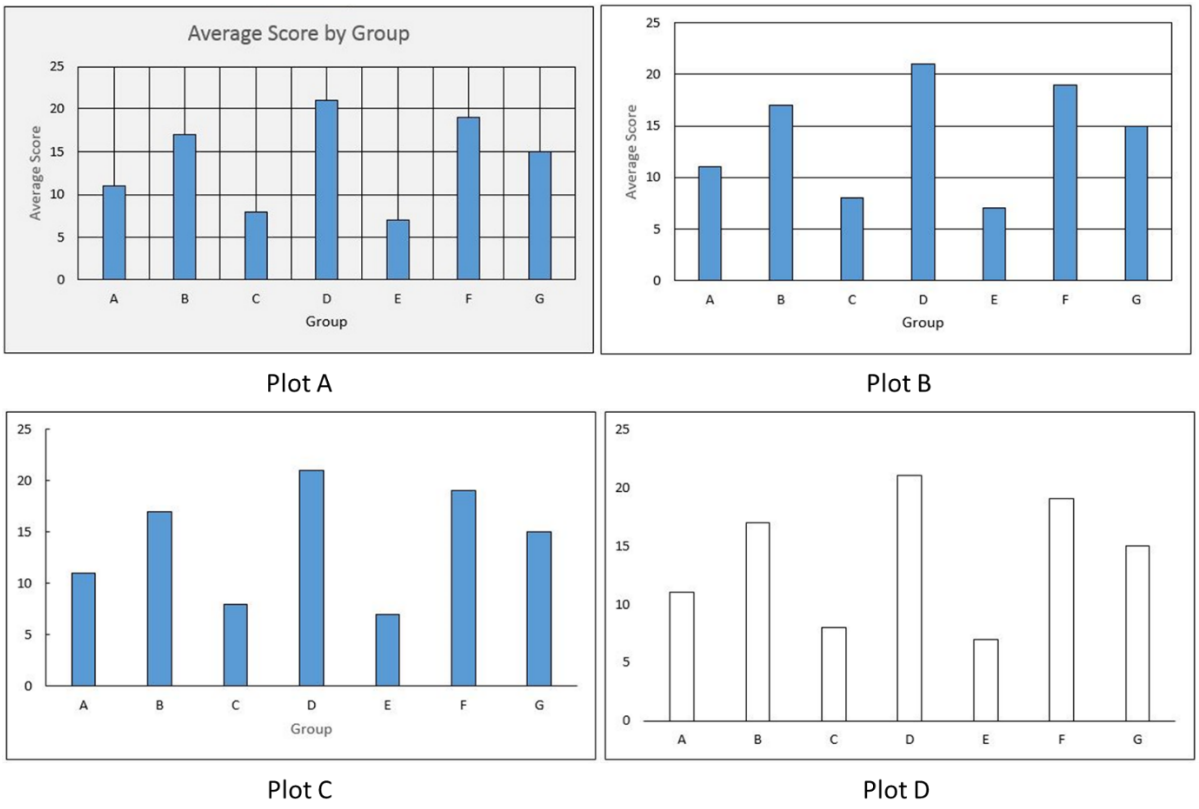


Figure 3: Barplots Presented to Survey Respondents

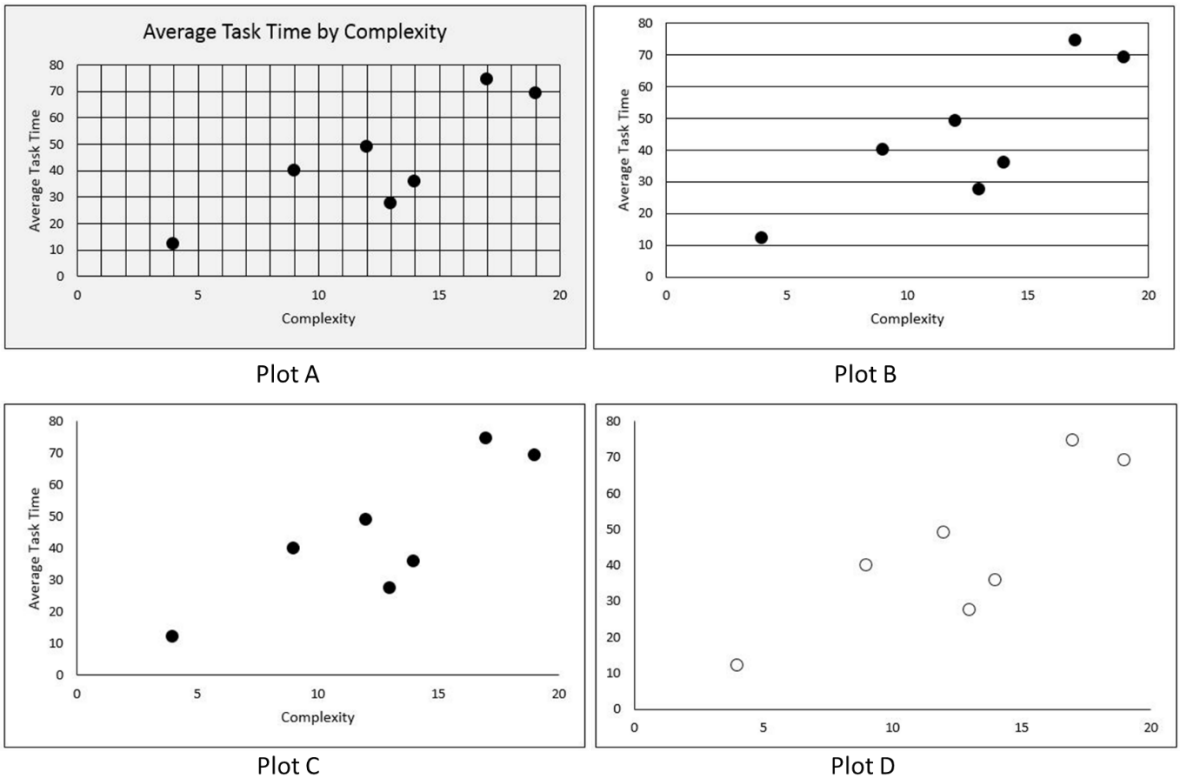


Figure 4: Scatterplots Presented to Survey Respondents

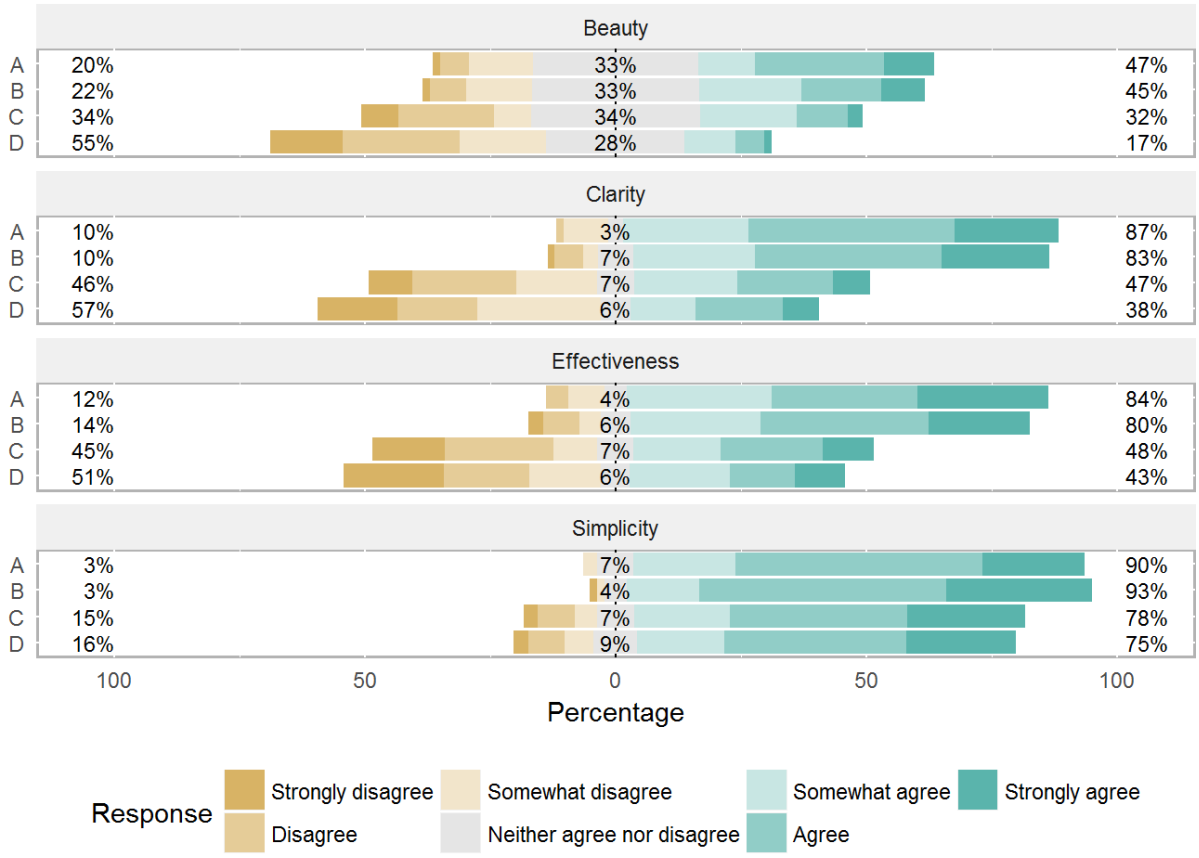


Figure 5: Student Survey Responses for Barplots A, B, C, and D (Percentage quantities are percentage of responses that were negative, neutral, and positive)

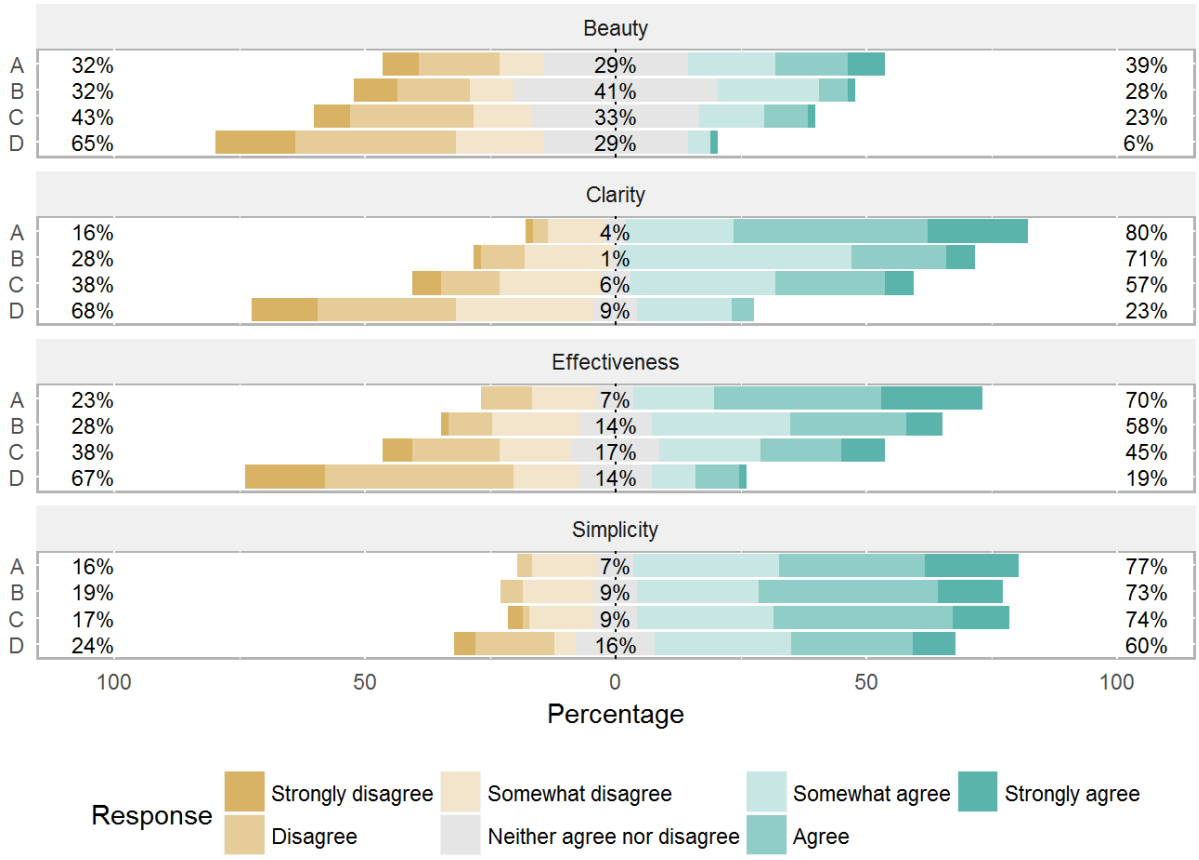


Figure 6: Student Survey Responses for Scatterplots A, B, C, and D (Percentage quantities are percentage of responses that were negative, neutral, and positive)

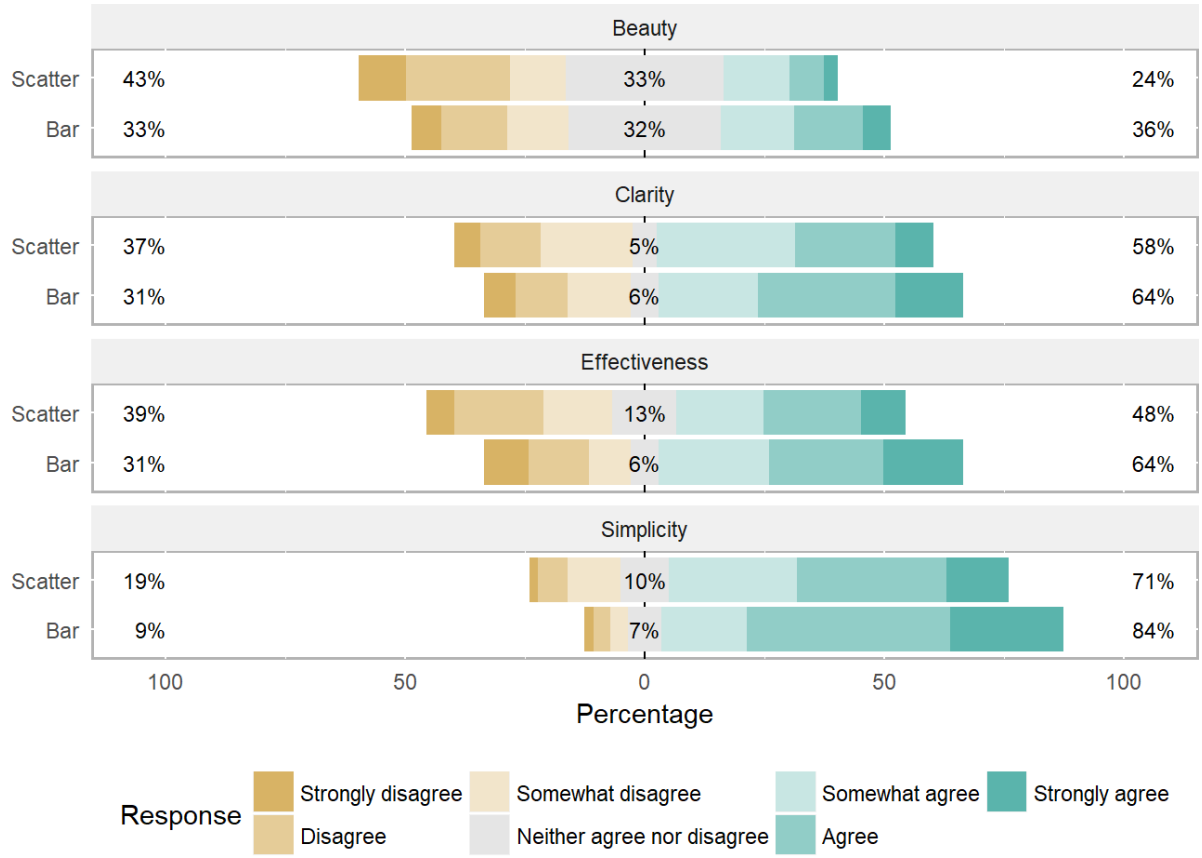


Figure 7: Student Survey Responses by Dimension, Grouped by Plot Type (Percentage quantities are percentage of responses that were negative, neutral, and positive)

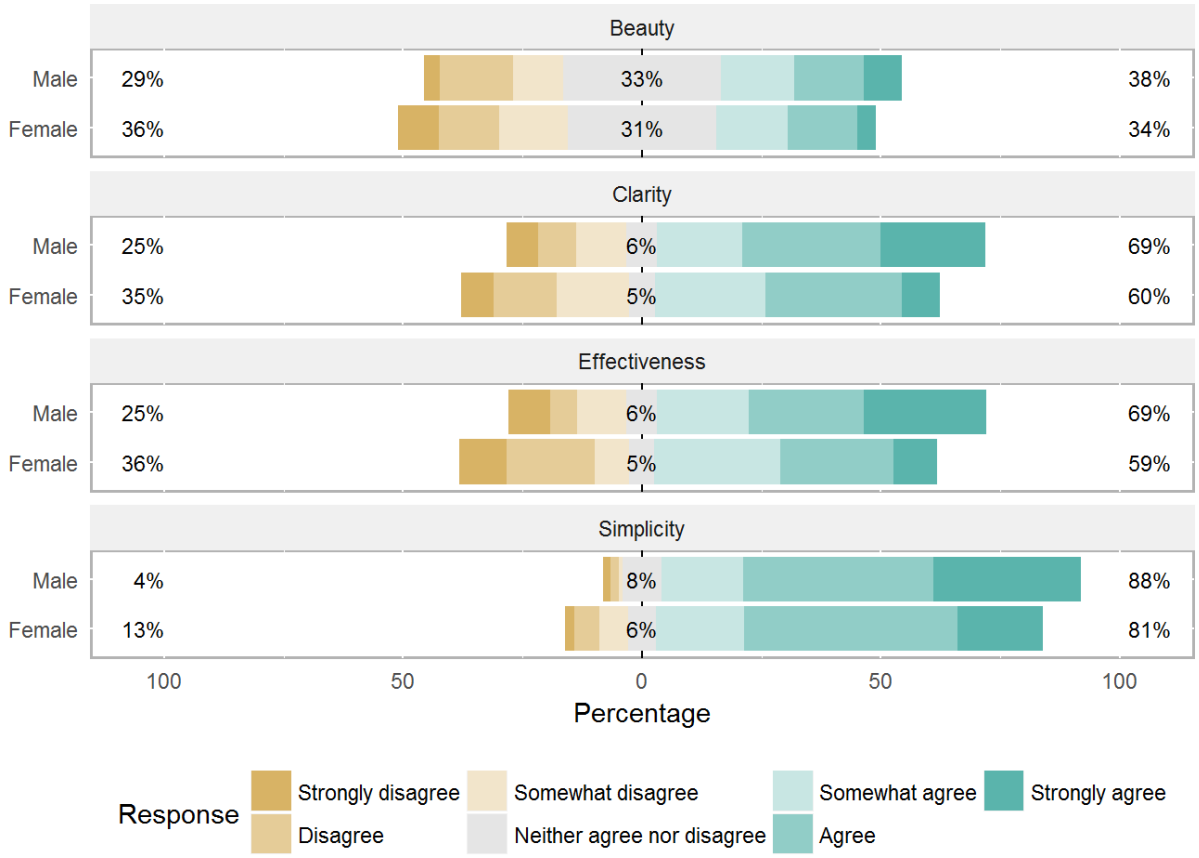


Figure 8: Student Survey Responses for All Barplots Grouped by Gender (Percentage quantities are percentage of responses that were negative, neutral, and positive)

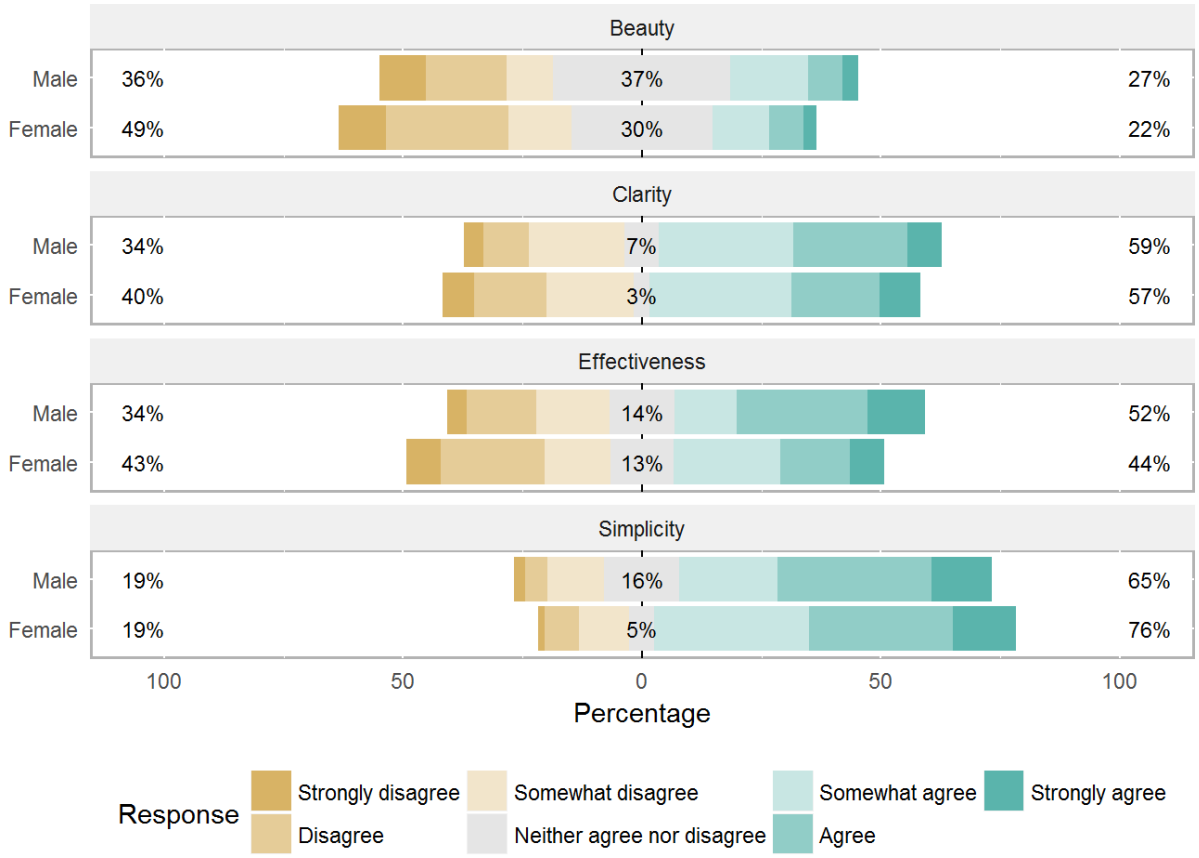


Figure 9: Student Survey Responses for All Scatterplots Grouped by Gender (Percentage quantities are percentage of responses that were negative, neutral, and positive)