The 11th Winona Computer Science Undergraduate Research Symposium

April 25, 2011

Winona State University
Winona, MN

Sponsored by the Department of Computer Science
at Winona State University

WINONA
STATE UNIVERSITY
Computer Science Department
http://cs.winona.edu

Table of Contents

A Model-driven User Interface in Predicting Limited-stage Small-cell Lung Cancer Survivability

Yingxu Liu, Mingrui Zhang
Computer Science Department, Winona State University
Winona, MN55987, USA
6127351042
Yliu09@winona.edu

ABSTRACT

The *Lung Cancer Survivability Prediction Tool* (LCSPT) is a web-based system, which gives doctors a statistical estimate on how long a patient may survive under certain treatments and health conditions. The current version of LCSPT contains three statistical models of *non-small cell lung cancer* (NSCLC). We integrate a *limited-stage small cell lung cancer* (LS-SCLC) model into the software tool. In doing so, we have researched a model-driven User Interface design and used it to design the user interface for four lung cancer treatment models. The model-driven design approach reduces the number of entries and avoids disordered entries on each interface, and it makes the LCSPT user friendly in a clinical environment.

Categories and Subject Descriptors

H.5.2 [User Interfaces]: Interaction style

General Terms

Design

Keywords

Model-driven user interface, model integration

1. INTRODUCTION

The *Lung Cancer Survivability Prediction Tool* (LCSPT) (Zhang et al., 2009) gives doctors a clear idea of how long a patient may survive given certain treatments and health conditions. The LCSPT was developed in 2006 under the guidance of the Computer Science Department at Winona State University and the Lung Cancer Study at the Mayo Clinic. The software tool is a clinical decision support system (CDSS). Currently it uses three statistical prediction models, two are *Non-Small Cell Lung Cancer* (NSCLC) models and the other is prognostic model for NSCLC Surgery. Both are developed in-house at Mayo Clinic in

Rochester. The first model predicts the patient's survival probability using only histological information including age, gender, stage, cell type, and tumor grade. The second model uses additional information, including the treatment options and the patient's smoking status. Both models were evaluated for their prediction accuracies on a test set of 1,518 patients (Sun et al., 2006). The evaluation was done by comparing the predicted and observed survival curves.

Proceedings of the 11th Winona Computer Science Undergraduate Research Seminar, April 25, 2011, Winona, MN, US.

Patients who underwent resection for primary non-small cell lung cancer use the third model. The factors associated with an impaired survival are sex, age, chronic obstructive pulmonary disease and so on.

The LCSPT is hosted on a computer server at the Mayo Clinic, and is accessible to both desktop computers and mobile devices. The supporting software as *Survival Probability Prediction Architecture* (SPPA) allows researchers to add and remove statistical models and to make changes to the inputs on the user interface. TABLE 1 summarizes the functions that can be performed by researchers, clinicians, and data entry people. The platform was designed for experimentation with diagnostic models and survival prediction. SPPA is based on the Model-View-Controller architectural pattern, as shown in Figure 1. It provides both a mechanism for defining models and a mechanism for testing the model in a clinical setting.

Table 1. Software Functions Provided to User Groups. (Zhang et al., 2009)

User Group	Functions Supported
Researchers	Add and remove prediction models in R; Change database for the model; Modify user interface
Clinicians	Add, view, and modify a patient record; Compare and select treatments
Data entry person	Add, view, and modify a patient record

The heart of the SPPA is the Controller. The controller was designed to be sufficiently general enough to enable quick and seamless modification of the system. The controller is subdivided into three components: the Model Manager, the View Manager, and the Variable Definition component. The Model Manager uses Java/R Interface (JRI) to provide an interface between the Java methods of the Controller and the prediction model, which is currently written in R programming environment. The View Manager is responsible for providing the researcher and/or the clinician with the results of the prediction model on a given patient. It consists of two components: the Web Form Generator and the Presentation Generator. Patient information is gathered through a web page form that is generated by the Web Form Generator. The glue that connects the Model with the View is the Variable Definition and Variable Mapping components.

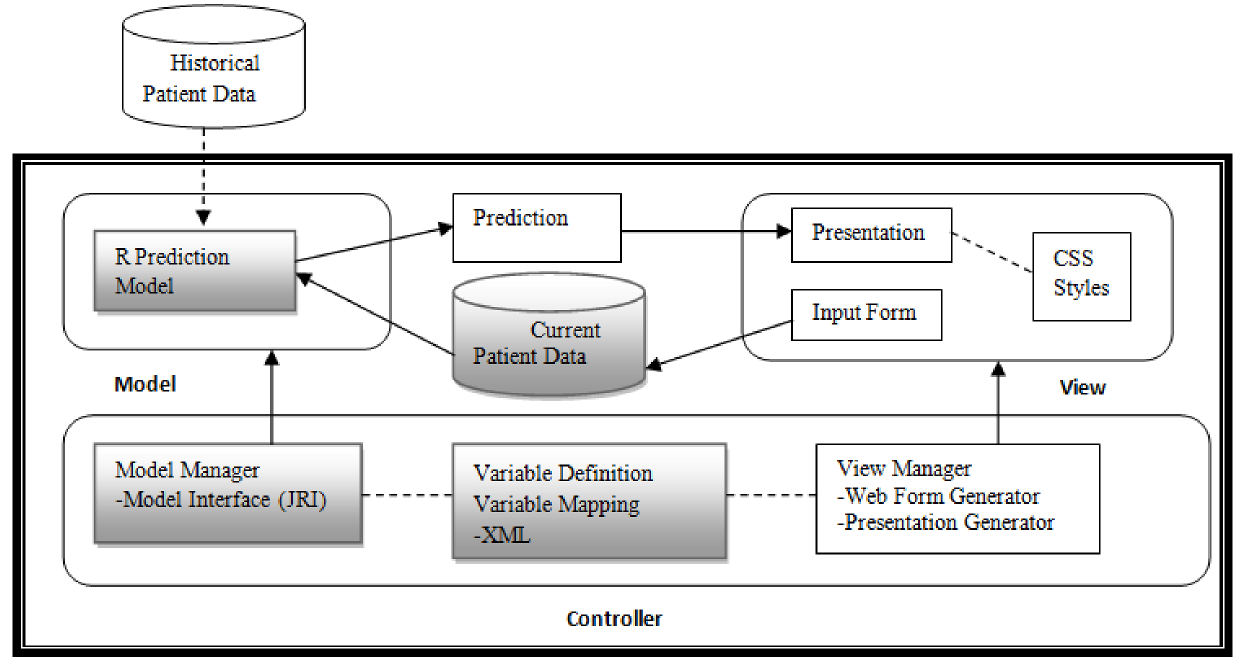


Figure 1. Survival probability prediction architecture, modified software units are highlighted. (Zhang et al., 2009)

The LCSPT is modified to include a limited-stage small cell lung cancer model. A recently published paper written by Dr. Jun Chen et al. (2009) shows that there is approximately 30-40% of small cell lung cancer (SCLC) belonging to limited-stage SCLC (LS-SCLC) at first clinical presentation. Based on the LS-SCLC statistical model which was developed at Mayo Clinic, the association of age at diagnosis, gender, years since quitting smoking, recurrence or progression, treatments and the patient's survival was indentified using the Kaplan-Meier method. A multivariable Cox proportional hazards model was applied to evaluate all of the above-mentioned variables for their independent predictive value on the patient's survival. This model was evaluated for the prediction accuracy on a test set of 284 patients (Chen et al., 2009). This finding provides useful information in treating LS-SCLC patients. However, the LS-SCLC model is not integrated in the LCSPT.

As more treatment models are added, the user interface entries could become cluttered and disordered. The situation becomes worse when the software application is used on mobile devices with small screens. If integrating the LS-SCLC model, with respect to original user interface design, the LCSPT will be difficult for doctors to use in a clinical setting.

2. HYPOTHESE

Model-driven User Interface design is more efficient in integrating multiple statistical models.

3. METHODS

In this project, we mainly focus on implementing and integrating the LS-SCLC statistical prediction model, design user interfaces for it, and assess their usability. Quantitative information is required to make a medical decision. Therefore, we developed model-based UIs to reduce inputs and avoid disordered elements per interface. Finally, the LCSPT has more comprehensive functions to process both the NSCLC and LS-SCLC statistical models.

3.1 **Database and XML Schema**

The lung cancer tool is made up of a web interface, database, and statistical models. The web-base tool is hosted on an Apache Tomcat server (http://tomcat.apache.org), and the patient's electronic records are stored on a mySQL server (http://dev.mysql.com) and processed via the R programming environment (http://www.r-project.org). Patient information is entered via a Web Interface, and a Java program uses JDBC to store data in the mySQL database (Figure 2). There are 34 tables included in the database, and each table has up to 175 variables. Java program retrieves patient information from the database and passes it to the statistical models. In the end, the R model returns a prediction to the Web Interface. We have designed several XML files to function as variable definition and mappings, the file cross references among Web Interface, database, and statistical models.

3.2 **Workflow**

Based on the structure of the LCSPT, we designed a workflow. The first phase was to develop variable definition and mapping files in XML. As the software units inside the LCSPT were developed independently, the data transfer is inconsistent and redundant. We developed XML mapping files to interpret variables which cross software units (Figure 3). For example, the "input" tags are used to map variables between the user interface and database; the "var" tags function as interpreting variables between the database and R statistical model; the “name”, “column”, “varname” and “item” tags refer to corresponding variable names in different software units; the “default” and “value” tags storing variable values are applied to data processing in R statistical model. The Web Form Generator uses the XML definition files to dynamically create the web form. Therefore, we designed definition files as normal and model-driven user interfaces (Figure 4) for later usability test. Normal user interface is a big screen with clustered variable entries. And, model-driven user interfaces are a sequence of user interfaces, which include a minimized number of inputs per screen.

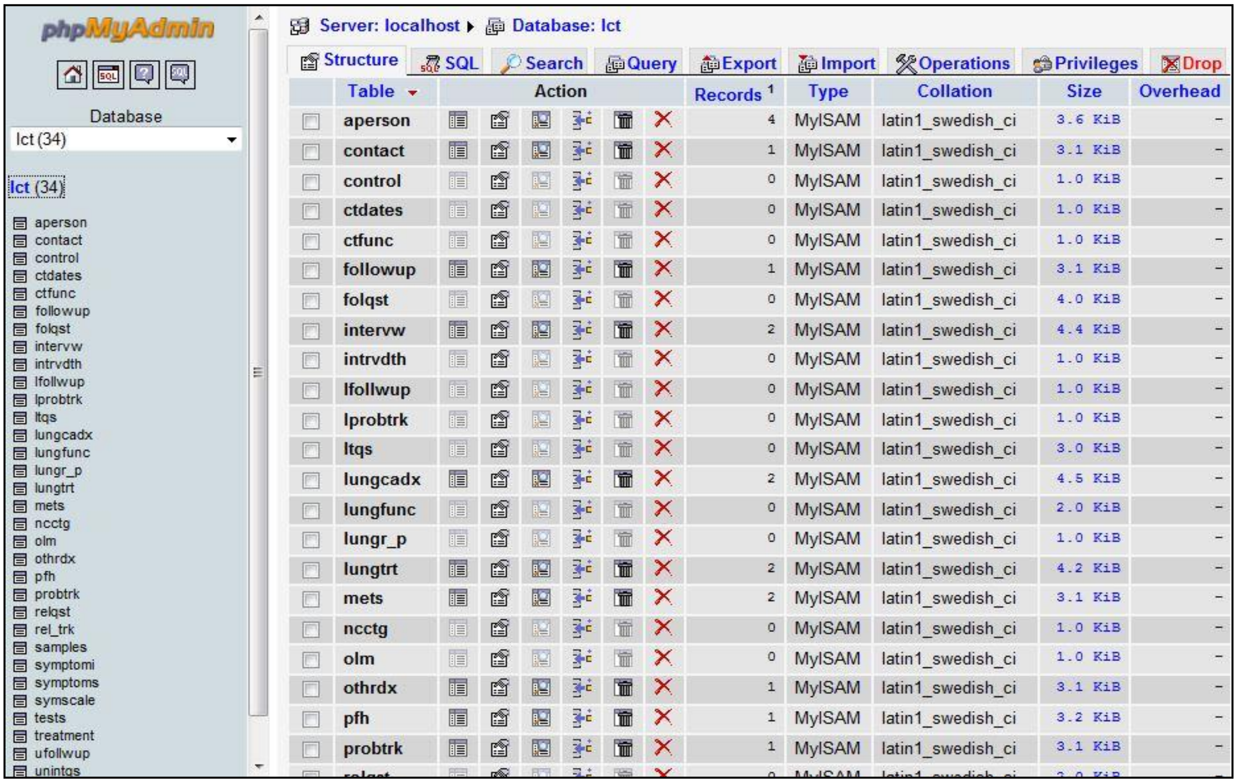
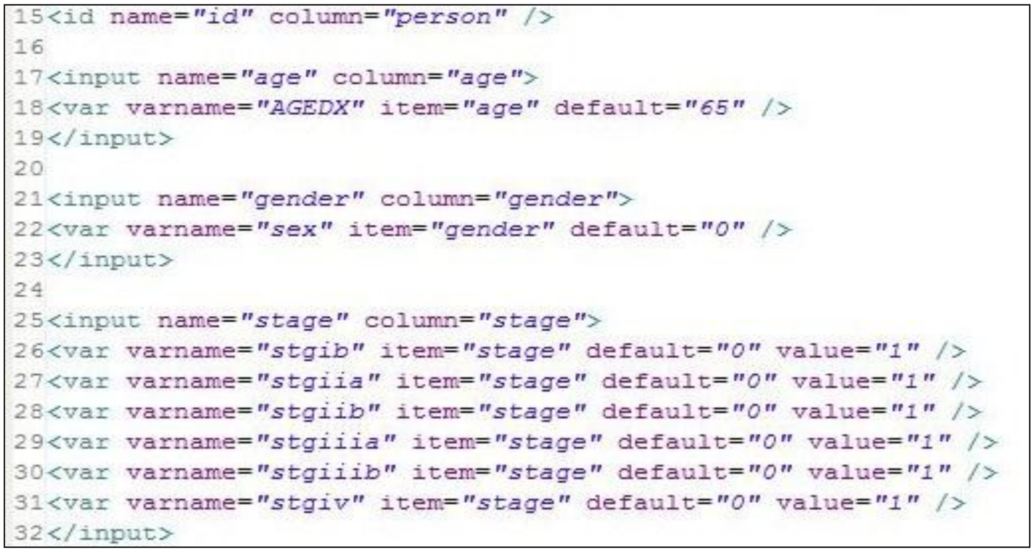


Table	Action	Records [1]	Type	Collation	Size	Overhead
aperson		4	MyISAM	latin1_swedish_ci	3.6 KiB	-
contact		1	MyISAM	latin1_swedish_ci	3.1 KiB	-
control		0	MyISAM	latin1_swedish_ci	1.0 KiB	-
ctdates		0	MyISAM	latin1_swedish_ci	1.0 KiB	-
ctfunc		0	MyISAM	latin1_swedish_ci	1.0 KiB	-
followup		1	MyISAM	latin1_swedish_ci	3.1 KiB	-
folqst		0	MyISAM	latin1_swedish_ci	4.0 KiB	-
intervw		2	MyISAM	latin1_swedish_ci	4.4 KiB	-
intrvdth		0	MyISAM	latin1_swedish_ci	1.0 KiB	-
lfollwup		0	MyISAM	latin1_swedish_ci	1.0 KiB	-
lprobtrk		0	MyISAM	latin1_swedish_ci	1.0 KiB	-
ltqs		0	MyISAM	latin1_swedish_ci	3.0 KiB	-
lungcadx		2	MyISAM	latin1_swedish_ci	4.5 KiB	-
lungfunc		0	MyISAM	latin1_swedish_ci	2.0 KiB	-
lungr_p		0	MyISAM	latin1_swedish_ci	1.0 KiB	-
lungtrt		2	MyISAM	latin1_swedish_ci	4.2 KiB	-
mets		2	MyISAM	latin1_swedish_ci	3.1 KiB	-
ncctg		0	MyISAM	latin1_swedish_ci	1.0 KiB	-
olm		0	MyISAM	latin1_swedish_ci	1.0 KiB	-
othrdx		1	MyISAM	latin1_swedish_ci	3.1 KiB	-
pfh		1	MyISAM	latin1_swedish_ci	3.2 KiB	-
probtrk		1	MyISAM	latin1_swedish_ci	3.1 KiB	-

Figure 2. Database tables



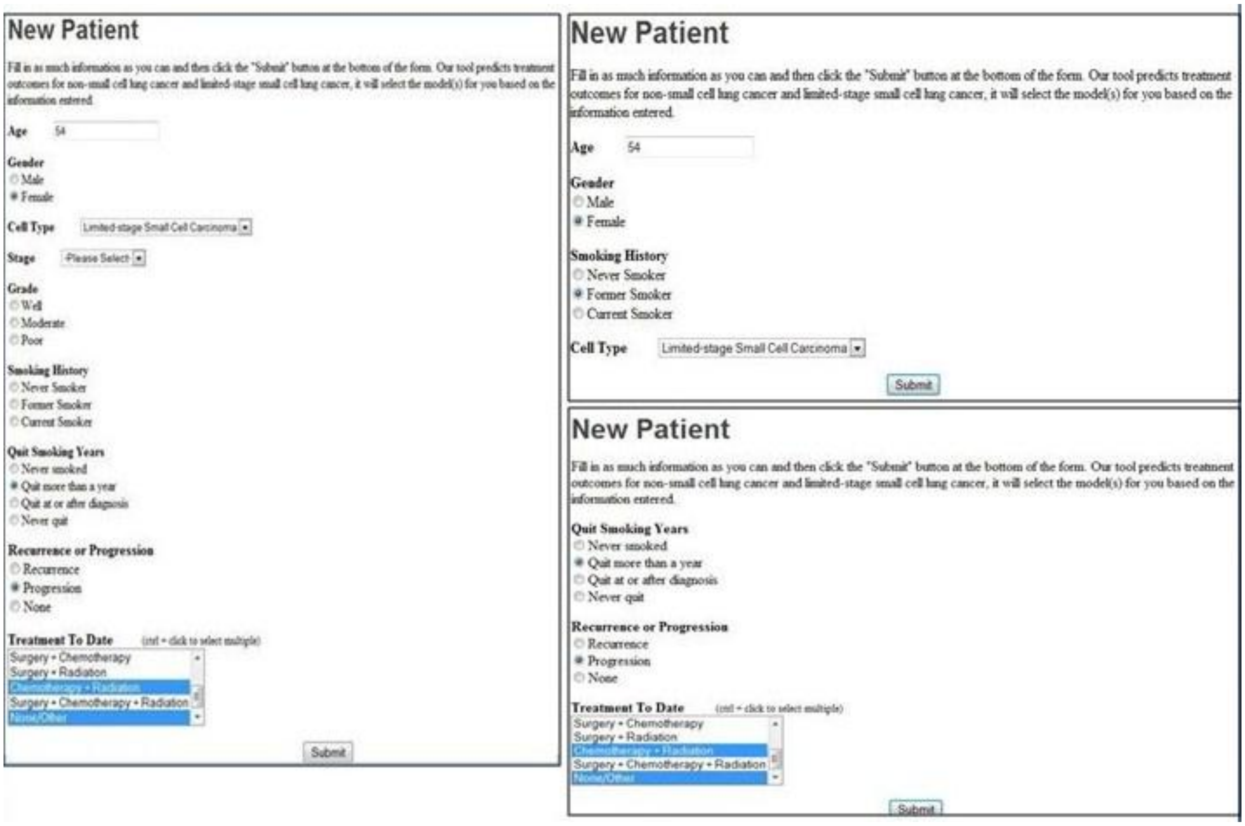
```
<id name="id" column="person" />

<input name="age" column="age">
<var varname="AGEDX" item="age" default="65" />
</input>

<input name="gender" column="gender">
<var varname="sex" item="gender" default="0" />
</input>

<input name="stage" column="stage">
<var varname="stgib" item="stage" default="0" value="1" />
<var varname="stgiia" item="stage" default="0" value="1" />
<var varname="stgiib" item="stage" default="0" value="1" />
<var varname="stgiiia" item="stage" default="0" value="1" />
<var varname="stgiiib" item="stage" default="0" value="1" />
<var varname="stgiv" item="stage" default="0" value="1" />
</input>
```

Figure 3. XML mapping file



New Patient

Fill in as much information as you can and then click the "Submit" button at the bottom of the form. Our tool predicts treatment outcomes for non-small cell lung cancer and limited-stage small cell lung cancer, it will select the model(s) for you based on the information entered.

Age 54

Gender
○ Male
● Female

Cell Type Limited-stage Small Cell Carcinoma ▾

Stage -Please Select- ▾

Grade
○ Well
○ Moderate
○ Poor

Smoking History
○ Never Smoker
○ Former Smoker
○ Current Smoker

Quit Smoking Years
○ Never smoked
● Quit more than a year
○ Quit at or after diagnosis
○ Never quit

Recurrence or Progression
○ Recurrence
● Progression
○ None

Treatment To Date (ctrl + click to select multiple)
Surgery + Chemotherapy
Surgery + Radiation
Chemotherapy + Radiation
Surgery + Chemotherapy + Radiation
None/Other

Submit

New Patient

Fill in as much information as you can and then click the "Submit" button at the bottom of the form. Our tool predicts treatment outcomes for non-small cell lung cancer and limited-stage small cell lung cancer, it will select the model(s) for you based on the information entered.

Age 54

Gender
○ Male
● Female

Smoking History
○ Never Smoker
● Former Smoker
○ Current Smoker

Cell Type Limited-stage Small Cell Carcinoma ▾

Submit

New Patient

Fill in as much information as you can and then click the "Submit" button at the bottom of the form. Our tool predicts treatment outcomes for non-small cell lung cancer and limited-stage small cell lung cancer, it will select the model(s) for you based on the information entered.

Quit Smoking Years
○ Never smoked
● Quit more than a year
○ Quit at or after diagnosis
○ Never quit

Recurrence or Progression
○ Recurrence
● Progression
○ None

Treatment To Date (ctrl + click to select multiple)
Surgery + Chemotherapy
Surgery + Radiation
Chemotherapy + Radiation
Surgery + Chemotherapy + Radiation
None/Other

Submit

Figure 4. The left and right are normal and model-driven user interfaces of the LS-SCLC

The second phase was to modify the mySQL database to store the information for the LS-SCLC patients, and change the Java program. To make the retrieval of patients' data convenient, addition variables were added to the existing database. According to the SPPA, we modified the Model component to integrate the LS-SCLC statistical model. Java code was implemented to invoke the new R model and display survival predictions. The LCSPT with normal and model-driven user interfaces are called LCSPT-1 and LCSPT-2.

Our concern with the model-driven user interfaces design is to limit the number of inputs per screen. Variable entries common to prediction models are placed on the interface that displays first, model specific entries later. A specific model would be picked by the system based on user inputs (Fig. 5). For instance, a user inputs age, gender, smoking history, and cancer cell type at the first user interface. If the user selects Limited-stage Small Cell for cell type, the LS-SCLC model is used in predicting the patient's treatment outcome. If the user selects others, his/her prediction model is based on the information of treatment type.

The prediction results from LCSPT consist of three tabbed pages, like the input of patient information, graphical (Figure 6) and tabular (Figure 7) views. In the graph view, the x-axis is the year of survival and the y-axis is the survival probability. By hovering over the cursor on the curve, the survival probability is displayed as an annotation. For instance, the annotation in the graph below indicates if under surgery treatment a patient's survivable prediction at 1.91 years is 60.06%.The tabular view lists the probability in specific years based on different treatments. To make the comparison of cancer treatments more convenient, checkboxes are used with the graph and table views of prediction results.

3.3 **Survey**

A usability test was based on the LCSPT-1 and LCSPT-2 version, which are the LCSPT with normal and model-driven user interfaces. We installed the two LCSPT in different desktops, and asked participants to use the system and finish a survey. The participants were ten students from Winona State University. Participants were divided into two groups, one group used the LCSPT-1 version first and then the LCSPT-2 version; the other one used the LCSPT-2 version then the LCSPT-1 version. They were asked to accomplish the task first then evaluate their experience on both versions of the LCSPT system. A patient information list was provided to each participant and let them process the same data. Participants were asked to create three new patients with specified, prognostic, and LS-SCLC information. Also, they were asked to find the survival possibility of a patient in 3.5 years from the charts shown on the page (Gegg-Harrison, Zhang, Nan, Sun, Yang, 2009).

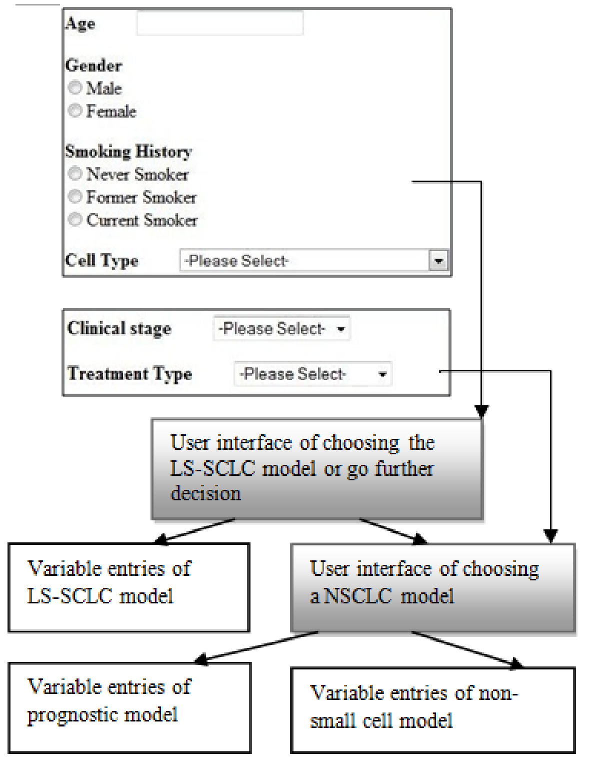

Age
Gender
○ Male
○ Female
Smoking History
○ Never Smoker
○ Former Smoker
○ Current Smoker
Cell Type
-Please Select-
Clinical stage
-Please Select-
Treatment Type
-Please Select-
User interface of choosing the LS-SCLC model or go further decision
Variable entries of LS-SCLC model
User interface of choosing a NSCLC model
Variable entries of prognostic model
Variable entries of non-small cell model

Figure 5. Data flow of the model-driven user interface design

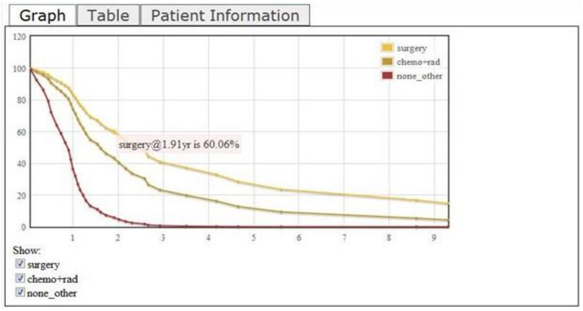
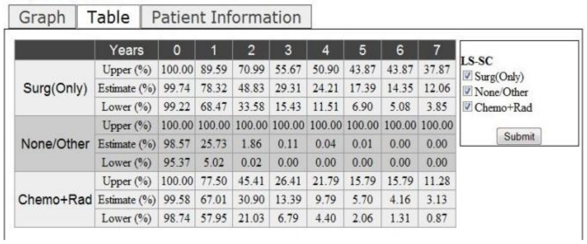


Figure 6. Graph view of prediction results from the LS-SCLC model



Graph | Table | Patient Information

	Years	0	1	2	3	4	5	6	7
Surg(Only)	Upper (%)	100.00	89.59	70.99	55.67	50.90	43.87	43.87	37.87
	Estimate (%)	99.74	78.32	48.83	29.31	24.21	17.39	14.35	12.06
	Lower (%)	99.22	68.47	33.58	15.43	11.51	6.90	5.08	3.85
None/Other	Upper (%)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	Estimate (%)	98.57	25.73	1.86	0.11	0.04	0.01	0.00	0.00
	Lower (%)	95.37	5.02	0.02	0.00	0.00	0.00	0.00	0.00
Chemo+Rad	Upper (%)	100.00	77.50	45.41	26.41	21.79	15.79	15.79	11.28
	Estimate (%)	99.58	67.01	30.90	13.39	9.79	5.70	4.16	3.13
	Lower (%)	98.74	57.95	21.03	6.79	4.40	2.06	1.31	0.87

LS-SC
☑ Surg(Only)
☑ None/Other
☑ Chemo+Rad

Submit

Figure 7. Table view of prediction results from the LS-SCLC model

After all the participants successfully finished the tasks they wrote down the survival possibility data they find from the charts on the LCSPT-1 and LCSPT-2 version. From Table 2, the mean is the time of creating new patients and finding survival possibilities. This data shows the LCSPT-2 version has a shorter finishing time and a smaller range. Therefore, in real medical situations, doctors would tend to save more time when using the LCSPT-2 version than the LCSPT-1 version. Furthermore, most of the participants were satisfied with the layout of the LCSPT-2 version. When asked about the ease with which to finish the tasks, participants all thought it was easy to accomplish in the LCSPT-2.

Table 2. Results

	Group 1		**Group 2**	
	LCSPT-1	**LCSPT-2**	**LCSPT-1**	**LCSPT-2**
Mean	6.333	4.333	4.667	4.333
Standard Deviation	1.247	0.471	0.471	0.471
Range	3	1	1	1

4. CONCLUSION

We developed model-driven user interfaces for the LCSPT in order to make it more efficient and convenient for doctors working in a clinical setting. The model-driven user interfaces successfully reduce inputs and avoid disordered elements per interface.

5. REFERENCES

[1] M. Zhang, S. Olson, J. Francioni, T. Gegg-Harrison, N. Meng, Z. Sun, and P. Yang, "Integrating R Models with Web Technologies," *International Conference on Health Informatics*, Porto, Portugal, 2009.

[2] Z. Sun, M. C. Aubry, C. Deschamps, R. S. Marks, S. H. Okuno, B. A. Williams, H. Sugimura, V. S. Pankratz, and P. Yang, "Histologic grade is an independent prognostic factor for survival in non-small cell lung cancer: an analysis of 5018 hospital- and 712 population-based cases," *J Thorac Cardiovasc Surg.*, vol. 131, pp. 1014-1020, 2006.

[3] J. Chen, R. Jiang, Y. I. Garces, A. Jatoi, S. M. Stoddard, Z. Sun, R. S.Marks, Y. Liu, P. Yang, “Prognostic factors for limited-stage small cell lung cancer: A study of 284 patients,” *2009 Elsevier Ireland Ltd*, 2009.

[4] T. Gegg-Harrison, M. Zhang, M. Nan, Z. Sun, P. Yang, “Porting a Cancer Treatment Prediction to a Mobile Device”2009.

An Evaluation of Skin Filtering in Face Detection

Brandon Hannasch
Winona State University Computer Science Department
Bhannasch07@winona.edu

Abstract

I have compared two face detection algorithms. Both algorithms use the Viola-Jones face detection, filters out non-skin tones from the search before running the algorithm. Both algorithms were run with a variety number of subjects per image to look at whether skin filter is more or less effective than the base algorithm depending on the number of subjects within the image.

Keywords

Face detection, Haar Cascade, Skin Filter, Viola Jones, Image processing

1. Introduction

The purpose of my research is to study and compare different algorithms designed to detect faces in images. Face detection is a important application of image processing. As stated by Abdallah S. Abdallah[1], detecting faces has important applications in commercial use, such as focusing before taking a picture, as well as being the important first step in face recognition and subject detection in security systems.

The most common type of algorithm uses pattern recognition to recognize different parts of a face, such as a nose or a mouth, and from there determine that a face is present. The primary type of face detection used in this research is the Viola-Jones method. The Viola Jones method uses a sliding window approach to find a face. This means that the algorithm tests for a face within small subsections of an image at a time.

As described by Robin Hastings [2], the Viola Jones method uses a Haar Cascade to determine whether or not there is a face within the window that algorithm is running. A Haar Cascade is a collection of Haar Features that define the structure being searched for in terms of light and dark patches.

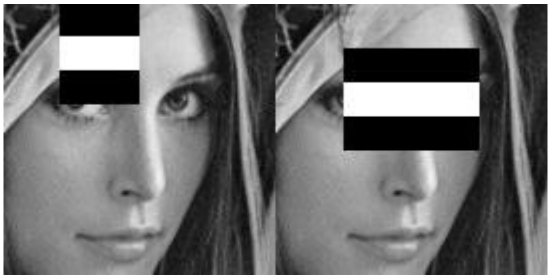


Figure 1 Haar Features

Proceedings of the 11th Winona Computer Science Undergraduate Research Seminar, April 25, 2011, Winona, MN, US.

As shown in Figure 1 above, a Haar feature for detecting eyes looks to find a rectangular area that is darker sandwiched between two lighter rectangular areas. Just because this pattern is recognized does not mean that the object is a face. Instead the surrounding area is also searched for other features, which, if found, give a strong possibility that it is a face. Figure 1 shows a small example of several of the Haar features that would need to be found in order for the face to be detected.

In order to be able to do these comparisons in nearly real time, the Viola-Jones method uses a technique known as Image Integration[1]. In image integration each pixel is assigned a value equal to all pixels between the pixel and a corner, typically all the pixels above or to the left of the pixels. Once image integration is completed finding the area of rectangles can be simplified to several addition and division equations rather than gathering data from all pixels in the region.

In addition to the Viola-Jones method one of the algorithms will use a skin filter. Skin filtering is a technique in which each pixel is checked through a database, and non-skin tones are denoted before the algorithm runs. When the sliding window moves over an area it first checks to see if the region is primarily skin tones. In not the process of using the Haar Cascade is skipped and the algorithm moves the window to the next area. According to Zaqout et al[3], by filtering out areas that definitely don't have face based on the skin color the algorithm can reduce the number of false positive results.

Since the same face detection algorithm is run over the same image in both approaches the major difference is that skin filtering reduces the area over which the algorithm searches for a face. Because of this, it's impossible for the algorithm to increase the number of positive face detections within an image. Instead this experiment focused on determining if the number of positive face detections decreased because the filter removed portions of faces in images. In addition, the number of incorrect detections, or false positives, was recorded to see how much of a positive impact the skin filter has.

Since the skin filter only runs if there is sufficient skin pixels to run, I also looked if an increase in subjects, which meant an increase of unfiltered subsections, had an impact on the images performance.

2. Hypothesis

An algorithm with skin filter will have fewer false positive results for images with different number of subjects without a significant increase in false negatives.

3. Methods

The images used in this study were be primarily taken from the "Faces in the Wild" face database set up by the University of Massachusetts[4]. The images in the database are compiled from news stories occurring in 2002 and 2003. The database was chosen because it contained a large number of images containing faces taken in a variety of settings. Each image in the database contains at least one subject and all images are sized between 105,000 and 180,000 pixels.

The images from the database were sorted based on the number of subjects that appear in the photographs. A subject was only counted if more than 50% of the person's face was visible (either both eyes or a majority of the face bellow the eyes were visible in the image). Eventually a dataset was randomly selected from the groups for each number of faces in an image 1-4, with at least 100 images in each set. Examples of the images can be found in the appendix.

The face detection software used in this experiment is called Faint[6]. Faint is a java plug in created by Malte Mathiszig using the Haar Cascade and basic algorithm from OpenCV[5] face detection library. All images were run through Faint twice; once with no extra filters and once with the skin filter.

An example of the output of Faint is shown below in figure 2. Faint denotes a found face on the image with a red rectangle around the region. Figure 2 is an example of the ideal output of the algorithm. All faces within the image have been marked and no marks appear where there is not a face.

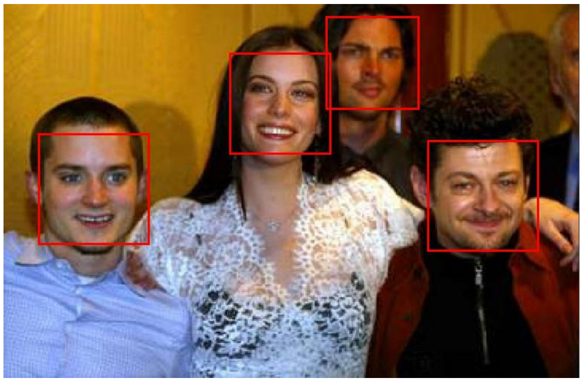


Figure 2 Example Output

The primary data collected from the study were the number of errors that the algorithms made. Errors were broken down into two categories. First, a face is present in the image but the algorithm does not detect it, known as a false negatives. An example of this is figure 3 where the ballplayer's face is present but the algorithm did not find it. False negatives were recorded as the percentage of faces missed.

The second category of errors is false positive results. In a false positive the system detects a face where no face is present. An example of this is figure 4 below. The algorithm correctly finds all three of the faces in the image but incorrectly finds a face on one of the subject's shoulders. False positives will be recorded in terms of the percentage of images with false positives.

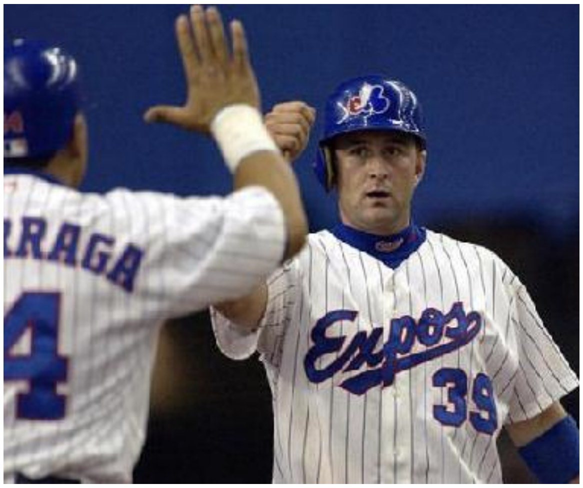


Figure 3 False Negative

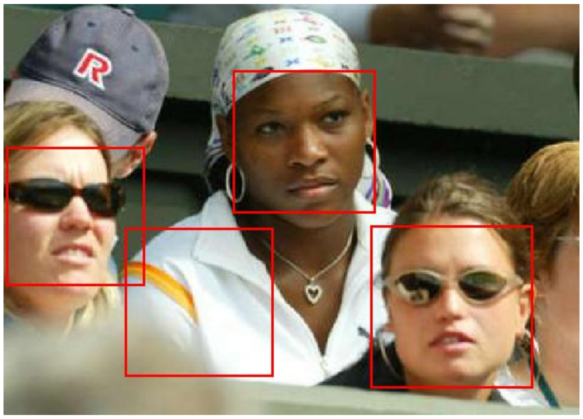


Figure 4 False Positive

4. Results

The results of the experiment are found in the following table.

Table 1 Results Table

	Percentage of Faces Missed		**Percentage of False Positives**	
# of Faces	Normal Algorithm	Skin Filter	Normal Algorithm	Skin Filter
1	7	7	19	9
2	16	16	40	26
3	15	15	38	22
4	16.75	16.75	36	30
Total	**15.1**	**15.1**	**33.25**	**21.75**

First, the percentage of faces found is exactly the same for both the algorithm with the skin filter and the one without. This seems to show that the skin filter in this implementation was done in such a way that no faces were filtered out. The algorithm with skin filter detects faces at an equal level to the non-filtering algorithm.

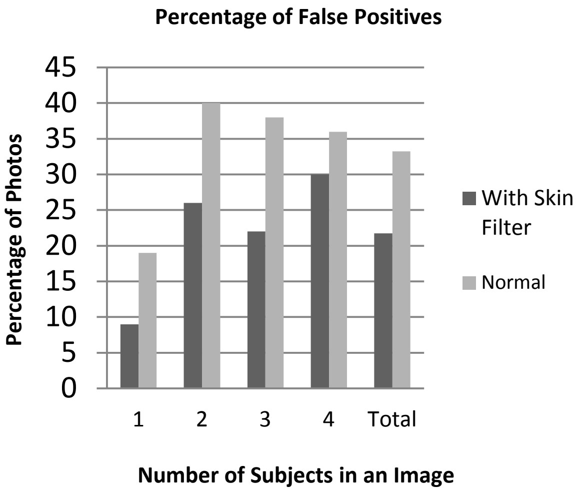


Figure 6 Number of False Positives

The above table is a visual representation of the number of false positives that both algorithms detected. The number of false positives found is less for all number of subjects in an image, with an average 11.5% decrease in the chance of getting a false positive. The decrease is the most drastic with one subject, removing more than half of the false positives found by the algorithm.

The reason the number of false positives did not drop further seems to be that the most common sites to get false positives are other areas on the body, such as necks, hands, and ears. Because these sites are skin colored they were not filtered out by the skin filter. This seems to be indicated at least in part by the increase of false positives with the addition of more than one subject. With the addition of more than one subject the camera cannot get a perfect angle on them all at the same time. This leads to more hands and sides of faces which in turn leads to more skin colored false positives.

5. Conclusion

In conclusion it does appear that a skin filter can reduce the amount of false positives that the Viola Jones algorithm detects by a significant amount without any reduction in face detection. The number of subjects in a photo does not affect the improvement from the skin filter, although having more than one subject does increase the probability of errors. It seems that further reduction of false positives would need to be focused on more accurate Haar cascades since many errors occur on other skin toned areas. In addition it would be interesting to see if the skin filter had an impact on the time that it takes to run the algorithm.

6. References

[1] Abdallah S., Abdallah “INVESTIGATION OF NEW TECHNIQUES FOR FACE DETECTION” Virginia State University (May 9th 2007)

[2] Hastings, Robin. “Seeing With OpenCV, Part 2: Finding Faces in Images” SERVO Magizine (February 2007) 48-60

[3] http://vis-www.cs.umass.edu/fddb/ accessed February 2nd 2011

[4] http://opencv.willowgarage.com/wiki/ accessed January 22nd 2011

[5] http://faint.sourceforge.net/ accessed February 4th 2011

[6] Zaqout, Ihab, Roziati Zianuddin, and Sapian Baba. 2004. "HUMAN FACE DETECTION IN COLOR IMAGES." *Advances in Complex Systems* 7, no. 3/4: 369-383. *Academic Search Premier*, EBSCO*host* (accessed January 20, 2011).

Appendix

The images below are samples from the images processed.

Online Authentication: Password Reuse and Security Questions

Jens Erickson
Winona State University
P.O Box 5838
175 West Mark Street
Winona, MN 55987
Jerickson07@winona.edu

ABSTRACT

Over the first decade of 2000, there was a major increase in the usage of online authentication using passwords for shopping, communication, gaming, and more recently, social media. The problem of password reuse, the usage of the same password for multiple accounts, increases the vulnerability of these accounts. Security questions are another method of security for online account. The rise of social media has changed what can be considered private information, bringing the reliability of these into question. These two related issues both show a human vulnerability in our online authentication. This study of WSU students takes an in-depth look at both of these vulnerabilities. It asks detailed questions about password reuse, as well as accessibility of information due to social networking. The collected data showed that the rate of password reuse from subjects was significantly higher than anticipated. Many subjects gave justification for this, but knew it was an issue. The results about security questions however showed that concern over security questions was lower than expected. The possible reasons behind these results and the implications of the results are discussed.

Categories and Subject Descriptors

K.6.5 [**Management of Computing and Information Systems**]: Security and Protection – *authentication, unauthorized access.*

General Terms

Measurement, Documentation, Security, Human Factors, Verification.

Keywords

password, password reuse, survey, security, security questions, authentication

1. INTRODUCTION

Online authentication is referring to the usage of a username and password to prove that the person who is using an account is the person they claim to be. It is not referring to the usage of biometrics or personal information like date of birth as ways to authenticate a person. Password reuse the growing problem of using the same password for multiple accounts, multiplying the odds of all of those accounts becoming compromised due to one becoming compromised. In late 2010, the passwords of over three quarters of a million accounts from Gawker, a parent site for many tech and media blogs, were compromised.[1] In early 2011, the passwords of around 61,000 accounts from rootkit.com were also compromised, being released in the public. A researcher from the University of Cambridge found that a intersection of the now public data revealed that within a sample size of 456 working email addresses, the password reuse rate was between 31% and 43%[2]. This goes to show the importance placed on unique passwords and the threat that password reuse provides. However, this is a large gap from earlier published works in password reuse. A user survey from Gaw and Felten in 2006 gave a result of a near 20% rate of password reuse[3]. Another empirical study from Herley and Florencio in 2007 had a reuse rate of around 12%. This gap shows that either one of two truths must hold, either the data from one of these sets is flawed, or the problem of password reuse has grown significantly over the last 5 years. In either case, new data must be collected.

Security questions are commonly used as a secondary way of accessing an online account or as a multi-factor identification tool. If a user cannot recall a password they set, they can answer a question they answered with private information when the account was created. However, the rise of social media has brought into question what types of information can be considered private. In Lori Kaufman's article, "How Private is the Internet" Lori goes into an analysis of how third parties track user web traffic, as well as discussion of privacy of emails and social networking. She states, "You shouldn't expect any privacy when you use the Internet" [4]. For a recent real life example, in 2008 the Yahoo email account of former vice-presidential candidate Sarah Palin was 'hacked'. The method of entry was answering a security question on the account asking where Palin and her husband met. This isn't just a problem for high-profile public figures though. In a symposium given by researchers from Carnegie Mellon University and Microsoft in 2009, resulted found that 28% of people who were close to the participants were able to correctly answer security questions.[5] This data shows that security questions are still an issue that needs to be looked at, more so with the advent of social media.

While these issues can look mundane, the implications of losing control of an online account extend farther than just identity on the internet. An estimated 11.7 million persons, representing 5% of all persons age 16 or older in the United States, experienced at least one type of identity theft in a 2-year period [6]. This means that awareness of these issues is quite important as fixing problems such as password reuse is extremely easy once the problem is identified. The aim of this study is to find just how prevalent the problem of password reuse is with undergraduate college students at Winona State University between the ages of 18 and 25. It also looks to see just how private information used to answer security questions is. Also, while previous studies have shown both password reuse and security questions to be a risk, the figures are out of date.

Proceedings of the 11thWinona Computer Science Undergraduate Research Seminar, April 25, 2011, Winona, MN, US.

2. HYPOTHESIS

This survey of Winona State University students will show two things. First, the rate of password reuse will exceed 60%, or five accounts per two unique passwords. Second, at least 30% or more of the surveyed students will state that they believe others could answer security questions they had answered.

3. OVERWIEW OF SURVEY

The method of data collection for this paper was a survey on the topic of passwords used for online accounts, ranging from games to bank logins, as well as finding the estimated number of online accounts the participant has. This survey included questions mainly about password reuse and security questions, but also included questions if the subject has ever had a compromised account, time using the Internet as well as social media, and password requirements of the accounts the subject has. All questions were looking for accounts, passwords, or security questions that the participant has used or encountered in the last 12 months. This survey was given to 37 participants. All participants were undergraduate students attending Winona State University and were between the ages of 18 and 23.

4. PASSWORD REUSE

Password reuse is difficult to quantify due to the number of accounts that a person has on the Internet. While previous surveys in this field have had exact figures on this subject, the number of accounts a person uses has grown significantly over the last 4 years, as our initial test surveys found. The initial survey asked the participants to give an exact figure on this question, but it was found difficult to provide an exact number in the testing stages of the survey. The fix was to have the participant recall within a range of accounts rather than an exact number. This gave a less accurate figure from the participant, but maintained the integrity of the data received from the participant. Out survey asked how many accounts a person had within ranges of 5, and how many passwords were used across these account. It again, also only asked for results from the previous 12 months, asking only to count passwords used in the last 12 months, and accounts accessed in that time period.

4.1 Methods

The survey began by asking giving the participant a list of 33 commonly used online sites organized into 6 differing categories (Blogs, Communication & Social Media, Commerce, Entertainment, Finance, and News.) Using this list, the participant was asked to recall as many online accounts that they had used in the last 12 months. The participant was then asked how many passwords were used between these accounts, again looking at the last 12 months. They then were asked a series of questions about relative strength of passwords used (from dictionary word to series of characters including special characters), reasons for password reuse or non-reuse, and awareness of password reuse. Later questions in the survey asked about online accounts that required password changes and differing requirements.

4.2 Results and Discussion

As stated before, we had 37 participants in this survey. All 37 of the participants were undergraduate students attending Winona State University. There were 16 male participants, and 21 female participants. Here in figure 1 is the breakdown of the participant's year of study.

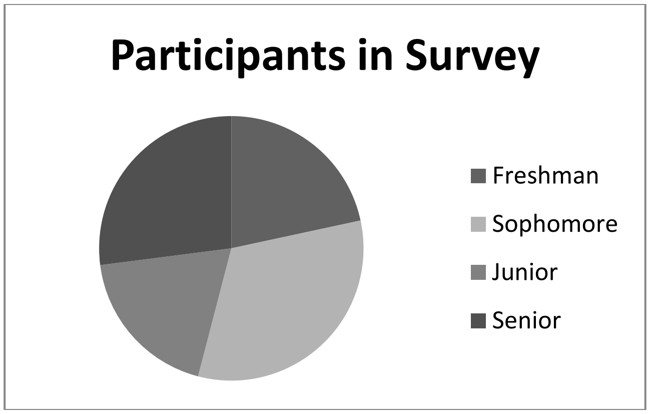


Figure 1. Participants year of study

The average number of accounts per participant was 18.8. The average number of accounts rose as the participant's years of study rose. Freshman had an average of 14.8 accounts, sophomores had an average of 16.5 accounts, juniors had an average of 20.5 accounts, and seniors had an average of 23.8 accounts. The average number of unique passwords is 5.2 per participant. By taking the number of accounts over the number of unique passwords a participant has, we get our password reuse rate (M=3.61, SD=1.85, Mdn=3.9). 26 of the 37 participants responded that they did use stronger passwords for accounts with access to financial information. The following question asked participants why they used passwords in the way they did, why they did or did not reuse passwords.

Table 1. Reasons for password reuse

26	Difficult to remember many passwords
6	Too many accounts to keep track of passwords
4	Same type of website, same password
1	Only have one password

Table 2. Reasons against password reuse

22	Restrictions between accounts differed
9	Security reasons
4	Account has privileged information
2	Different type of website

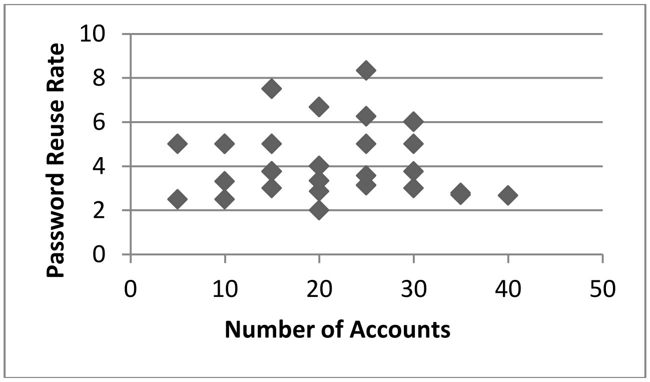


Figure 2. Rate of Password Reuse

There are a few implications that we can read out of this. The biggest is looking at figure 2, we see that rather than the rate of password reuse rising as the number of accounts a participant had rose, it stays the same, if not lower than with less accounts. The reason for this is in Table 2, restrictions on passwords differ on accounts. This type of biodiversity is a natural and unintrusive way to deal with password reuse that is in place almost without our knowledge. There was a distinct link between a participant's number of accounts and the level of security they put into accounts. 22 of the 26 participants who used tiered levels of passwords had over 16 accounts (out of 24 respondents in that category). This, along with data from the second portion, suggests that people who spend more time online are informed of threats to those accounts. Another development that was in line with previous data was the number of accounts averaged by different years of study rising over the years. As time goes on, we only accumulate more accounts, forgetting less than we gain, with the ones we do forget still existing.

5. SECURITY QUESTIONS

The security question section of the survey had 3 goals that we wanted to find. First was how many of the participants believed that others would be able to answer these questions, with respect to how much of that information would be available. Second is what the participant's presence on social media sites was. Third was to find out which questions were most commonly used by the participants.

5.1 Methods

The participant was provided with regularly used security questions from sites ranked in the Alexa Top 100. They were asked if they answered questions of this type before and if they could recall any specific questions. They then were asked a series of questions to answer on a 5 point scale about the level of security these questions provide, the ease of answering these questions themselves, and by others. They also were asked about any social media presence they had.

5.2 Results and Discussion

Participants were asked about what security questions they had been asked in creation of accounts they had. 22 of the 37 participants stated that when asked, they when given the chance, they would create a user-generated question over the given choices. Figure 3 shows the results when participants were asked about how they felt other users would be able to answer security questions they used.

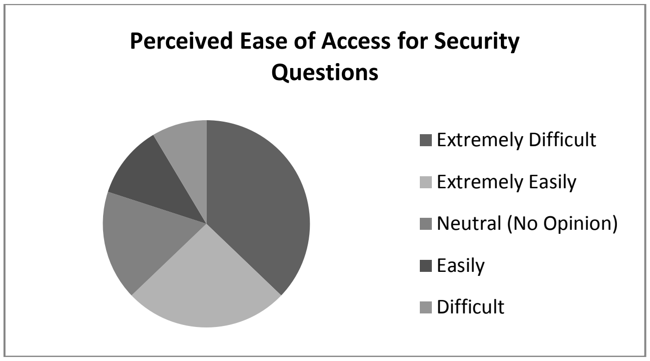


Figure 3. Ease of Access

When asked about the level of security that these questions provided, 7 felt an account with a security questions was more secure, 11 felt there wasn't a difference with or without, and 19 felt less secure using them. The participants were also asked about any social media presence they had, such as a Facebook or Twitter account, to which 35 of the 37 participants stated they did have. Out of our 37 participants, 3 had a compromised account within the last 12 months. As stated before, the trend of users with more accounts having a more informed level of threats continues. Those participants with more accounts also showed more concern with security questions. Out of the 9 people who felt a security question could be answered with extreme ease, 6 answered they had 21 or more online accounts, with 8 having more than 16. The most interesting result gathered from this portion is the level of security that security questions provide. Whether it is due to popular culture or the amount of incidents where security questions are used as a back door, participants felt they were more of a liability.

6. CONCLUSIONS

From the collected data, we can see that password reuse is a growing problem, concurring with the conclusions that were reached in Gaw and Felten's work.[3] However, while they correctly predicted the rise of accounts we would have in the future, they did not predict the measures that were already taken to stem the tide of password reuse. There are many measures that can be or already are being taken to end the problem of password reuse. The biodiversity of password requirements looks to be a very good first step in stopping password reuse, while being quite non-intrusive to the user, unlike such measures as aging passwords. Another way to solve the problem of password reuse is the consolidation of Internet accounts. Both Facebook and Google accounts can be used as valid logins for many different online websites, removing the need for having many different accounts for these sites. This however introduces a single point of failure which is easier to take advantage of than password reuse does. It is a step forward for convenience but a step backward as far as security goes, solving a problem while introducing another. The other major solution to the issue of password reuse is multi-factor authentication. Many banks already implement this, as well as a select few online services that require this. By having additional login requirements outside of just a password, the impact of having a password compromised is diminished, though not removed.

Security questions pose a different problem, one that our participants did not seem to be as aware of as we had hoped, though they were keenly aware that security questions can pose another threat to an online account. The solutions to the problem of security questions lie in the information asked by them, and the users answering them. Questions posed by these security questions need to move to older information that wouldn't be accessible by usage of social networking. On the other hand, users need to be careful of what information they choose to reveal online, as once information is revealed, it is extremely difficult if not impossible to remove that information from the Internet. A good solution that is implemented by every site we used in the survey is allowing the user to write-in their own question, removing the predictability of the questions.

6.1 Future Possibilities of Exploration

This field is one that has a variety of directions that all need exploring in the future to ascertain the level of which issues such as password reuse effect people. A larger dataset would be a great area to move to next on this topic. Another interesting direction to take both of these issues would be to look at a different age group all together. Rather than looking at college students between the ages of 18 to 23, a look at these habits on a group of middle age workers between the ages of 35-45 would provide extremely different results on all levels. For this survey, we instructed the participant to treat passwords with minor alterations such as the addition of a number at the end or altering case as different passwords. However, we did not see how predominate these practices actually are, or the risk that they pose. While password reuse is one way to look at the issue of shared authentication between online logins, another would be to take a look at username reuse, as they are the other key component to online authentication.

7. REFERENCES

[1] Bonneau, Joesph. The Gawker hack: how a million passwords were lost Light Blue Touchpaper, University of Cambridge, 2010.

[2] Bonneau, Joesph. Measuring password re-use empirically Light Blue Touchpaper, University of Cambridge, 2011.

[3] Gaw, Shirley; Felten, Edward. Password management strategies for online accounts ACM Digital Library, Princeton University, 2006.

[4] Caufman, Lori. How Private is the Internet? IEEE Security and Privacy Jan/Feb 2011, Carnegie Mellon University

[5] Schechter, Stuart; Brush, A.J.; Egelman, Serge. It’s No Secret. Measuring the Security and Reliability of Authentication via ‘Secret’ Questions, 30th IEEE Symposium on Security and Privacy, Microsoft/ Carnegie Mellon University 2009

[6] Langton, Lynn and Planty, Michael. Victims of Identity Theft Bureau of Justice Statistics, 2008