# The Knot or The Noose?
# Analysis of Privacy on a Wedding Planning Website

KATIE A. SIEK

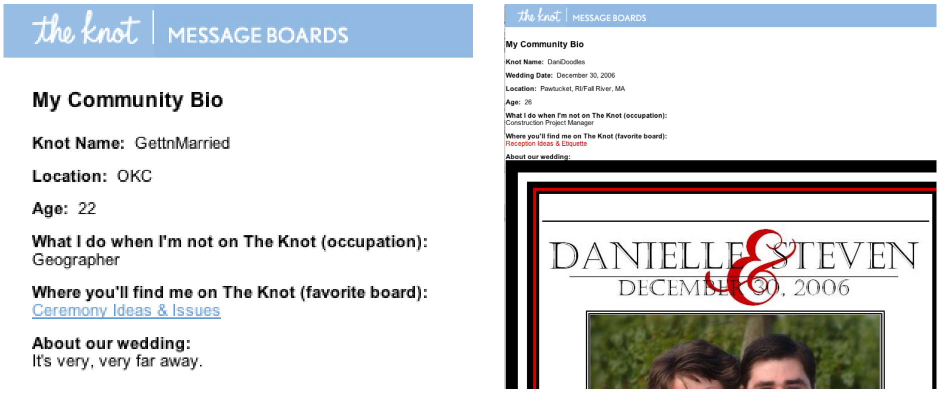
**Abstract**

At any given time, at least 2.4 million couples are planning their weddings. Couples get advice from friends and family, books, magazines, television, and the Internet to help plan their weddings. The Knot, an online wedding planning resource, helps couples plan and brag about their weddings with personalized online web-pages called bios. The amount of information varies in each bio, however some bios have enough information to help malicious third parties create phishing schemes, identity thefts, cancellation problems, and robberies. This paper presents a statistical analysis of privacy concerns for couples on the Knot and shows the Knot is sometimes a noose for unsuspecting couples who divulge too much information.

## 1 Introduction

The Knot, a publicly traded wedding media and services company, is known for a monthly magazine and regularly updated web site. The web site assists newly engaged couples gather information about wedding etiquette and services for their upcoming nuptials. The Knot web site, here on referred to as The Knot, has everything from what kind of dresses are available in the winter to an online store where couples can buy personalized fushia matchbooks. The site boasts 2.1 million unique visitors a month and at least 3,600 couples joining The Knot every day [1]!

The "talk" section of The Knot is a popular web board where members can ask questions they would not ordinarily feel comfortable asking to friends and relatives with the feeling of anonymity. Questions range from information about dresses to envelope etiquette. For example:



**My Community Bio**

**Knot Name:** GettnMarried

**Location:** OKC

**Age:** 22

**What I do when I'm not on The Knot (occupation):**
Geographer

**Where you'll find me on The Knot (favorite board):**
Ceremony Ideas & Issues

**About our wedding:**
It's very, very far away.

(a)

**My Community Bio**

**Knot Name:** DaniDoodles

**Wedding Date:** December 30, 2006

**Location:** Pawtucket, RI/Fall River, MA

**Age:** 26

**What I do when I'm not on The Knot (occupation):**
Construction Project Manager

**Where you'll find me on The Knot (favorite board):**
Reception Ideas & Etiquette

**About our wedding:**

(b)

Figure 1: (a) Bio with little information; (b) Bio with detailed personal and wedding information

> My fiance thinks its okay for the groomsmen to wear a mix of shawl and peak tuxedos. I want everyone to MATCH. All of my bridesmaids *bought* dresses to match. The least the guys could do is *rent* matching tuxedos! Am I a bridezilla?

---

> Do you think my Jim Hjelm dresses Style JH0055T in Tranquil and Fawn (pic in bio) will look okay with calla lilies dyed raisin? My mother thinks it will be hideous. But it's my wedding. I like it. What do you think?

The last quote from The Knot web board mentions the member's bio (short for biography). A bio is a web page that has the member's user-name and other information about the member or member's wedding plans. Some bios only have the member's name, while others have detailed personal and wedding information, as shown in Figure 1.

The personal information available in some bios could leave Knot members vulnerable to phishing scams, identity theft, cancellation problems (e.g. canceling the wedding and/or vendor(s)), and robbery. In this paper,I examine how much personal information Knot members make publicly available and how much personal information can lead to privacy concerns.

I begin with a review of privacy concerns. The experimental design is discussed in Section 3. In Section 4, I describe the data analysis in detail. My findings are in Section 5. I conclude with a discussion of the results and future work.

## 2 Review of Privacy Concerns

We have all received email spam about millions of dollars wrapped up in a Nigeria bank account and how we can get a share of the money if we simply send our bank account information to some email address [6]. Most of use would never dream of emailing a stranger our bank account information. We ignore junk email, put anti-virus on our computers, and look for the little lock icon when we make purchases online in hopes that all of this is enough to keep our "digital self" safe from privacy and security schemes. However for those of us who have an on-line presence, like Knot members who create biographies, we must take extra care to ensure we do not readily make available too much personal information. In this section, I will discuss privacy concerns that may happen to knot members who post too much personal information about themselves on the Knot web board and personal bio.

In a *phishing* attack, a malicious third party will deceive a victim to get secret information such as passwords, bank account numbers, payments, etc. The Nigerian email scheme is an example of a phishing attack. A context aware phishing scheme attempts to make the victim feels comfortable with the authenticity of the email message received by using timing and context [4]. For example, if a Knot member had the vendor names and her email address (context) in her bio or on a web board, a malicious third party could create an authentic looking invoice and email the Knot member asking for payment to a Paypal account. Since the member is on The Knot, it is assumed the wedding is in the near future or happened recently. Thus receiving invoices from vendors would not seem suspicious (timing) and the malicious third party could steal money from the member.

*Identity theft* happens when a malicious third party uses a victim's personal information, such as one's name, Social Security number, credit card number, or other personal information, without the permission of the victim to commit fraud or crime(s) [3]. In previous research, my colleagues and I found that with just a name or email address, we could get enough information to create an identity theft scheme [2]. Thus, if a Knot member had her email address or full name posted in her bio or on the web board, we could get enough information to steal her identity.

*Cancellation problems* happen when a malicious third party finds out a future couple's wedding vendor(s) and/or venue information and pretends to be the couple or wedding coordinator. The

| **Privacy Concern** | **Information Needed** | **Helper Information** |
|---|---|---|
| Phishing | Email and Vendor Information | Online Payment |
| Cancellation Problems | Vendor Information, Brides and Grooms First name, Date of Wedding | Time of Wedding, Bride and Grooms Last Names |
| Identity Theft | Email or (Full name of Bride or Groom, Age, and Hometown) | Occupation, Hosting pictures on Wedding or Personal web site, and personal picture |
| Robbery | Full Name of Bride or Groom, Hometown, and Date of Wedding | Occupation, Hosting pictures on Personal Site, and pictures of home |

Table 1: Summary of privacy concerns and bio information needed to implement the scheme

malicious third party will contact the wedding vendor(s) and/or venue and cancel the wedding. Cancellation problems are occasionally discussed on The Knot when a Knot member is contacted by the wedding venue or vendor(s) to confirm her wedding cancellation.

The last privacy concern is *robbery.* I define robbery as the act when a malicious third party breaks into a victims home and steals belongings. Knot members sometimes have wedding and honeymoon dates in their bios. If the member's address can be established and the dates of the wedding and/or honeymoon are known, robbery can occur.

People with no privacy concerns are defined as those who do not have any names, hometown(s), email, vendor or venue information posted and host pictures on a picture hosting site that is not directly related to the individual (such as Yahoo or Snapfish).

All of these privacy concerns should be kept in mind when I discuss how data was collected and what variables I looked at in the Knot bios.

## 3 Experimental Design

Original data was collected via a survey shown in Appendix A. Each question in the survey corresponded to information needed to implement a privacy attack as shown in Table 1. "Information Needed" refers to the minimum amount of information needed to implement the privacy attack. "Helper Information" refers to information that is not necessary to implement the attack, but could assist in creating a more realistic attack. All of the information needed for these attacks could be available if the bio pictures are hosted on a personal web site. A malicious third party could look at the bio, view the picture independently (i.e. right click on the image), and then see what other information is available on the personal web site.

A Human Subject Committee (HSC) approved call for participation notice was posted on all of the Knot web boards at various times of the day. Interestingly enough, I had to put my full name and university contact information (including email) on the call for participation for this *privacy* study.A total of 262 Knot members participated in the survey.

After looking at the initial data and possible privacy concerns, I decided to create a study analyzing the privacy of members on The Knot. More specifically, may goals were to find out:

- How much personal information do Knot members make publicly available?
- How much of the personal information could lead to privacy concerns?

# 4 Data Analysis

Before I analyzed the data, I created twenty-nine variables listed in Appendix A based on the survey. The variables were binary coded - a 1 signified the variable was in the bio and a 0 signified the variable was not in the bio. When users filled out the survey, the PHP backend would create tab delimited strings of 1s and 0s based on the Knot members' input and email the results to the researchers. The strings of 1s and 0s were saved into a text file and read into R, a statistical computing and graphics programming language environment [5], as a binary incidence matrix.

Trying to see trends in the raw binary data of 262 cases (Knot member responses) and twenty-nine variables was difficult, if not impossible. I needed a way to identify patterns and summarize similarities and differences. Exploratory multivariate analysis, more specifically Principle Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA), was the ideal way to analyze the data in R.

PCA (`princomp`) was performed using LAPACK routines to produce eigenvalues and eigenvectors (`eigen`) on the covariance matrix. The eigenvectors are the principle components - they show relationships of how the data is characterized together. The eigenvalues assist in recalculating the original data set if necessary and ordering the principle components from highest to lowest significance [7]. For our analysis, I plotted the results from the principle components (shown in Appendix B and decided to use the first four principle components. Since I only took four out of the twenty-nine principle components (one principle component per variable), I lost some data but the principle components/variance is small and insignificant.

The R command, `princomp` returns (among other things) a matrix of principle components called `loadings` and the `scores` of the original data when the principle components are applied to the incidence matrix. I plotted the matrix of principle components (`loadings`) on the four selected principle components to help us understand the relationships between the survey variables and privacy concerns. The variables in the plot were color coded based on a frequency scale of 1 to 5, where 1 meant the variable was rarely true and 5 meant the variable was true for most of the Knot members who completed the survey. I added lines and circles to each loadings plot to indicate the origin in each plot and what variables are significant.

Plotting the principle component scores by themselves is not meaningful because there is no clear way to delineate clusters of Knot members.I used HCA to delineate clusters by first creating a distance matrix (`dist`) using the scores from the first four principle components. Once I had the distance matrix, I used HCA with Wards method (`hclust`) to create compact, spherical clusters. Then, we plotted the result from the clustering analysis and analyzed the dendrogram shown in Appendix B. I chose to use a cut of three clusters because they had the most variation between tree branches and could be easily identified when color coding the PCA scores plotted on the first four principle components.

In the next section I will discuss each plot in more detail.

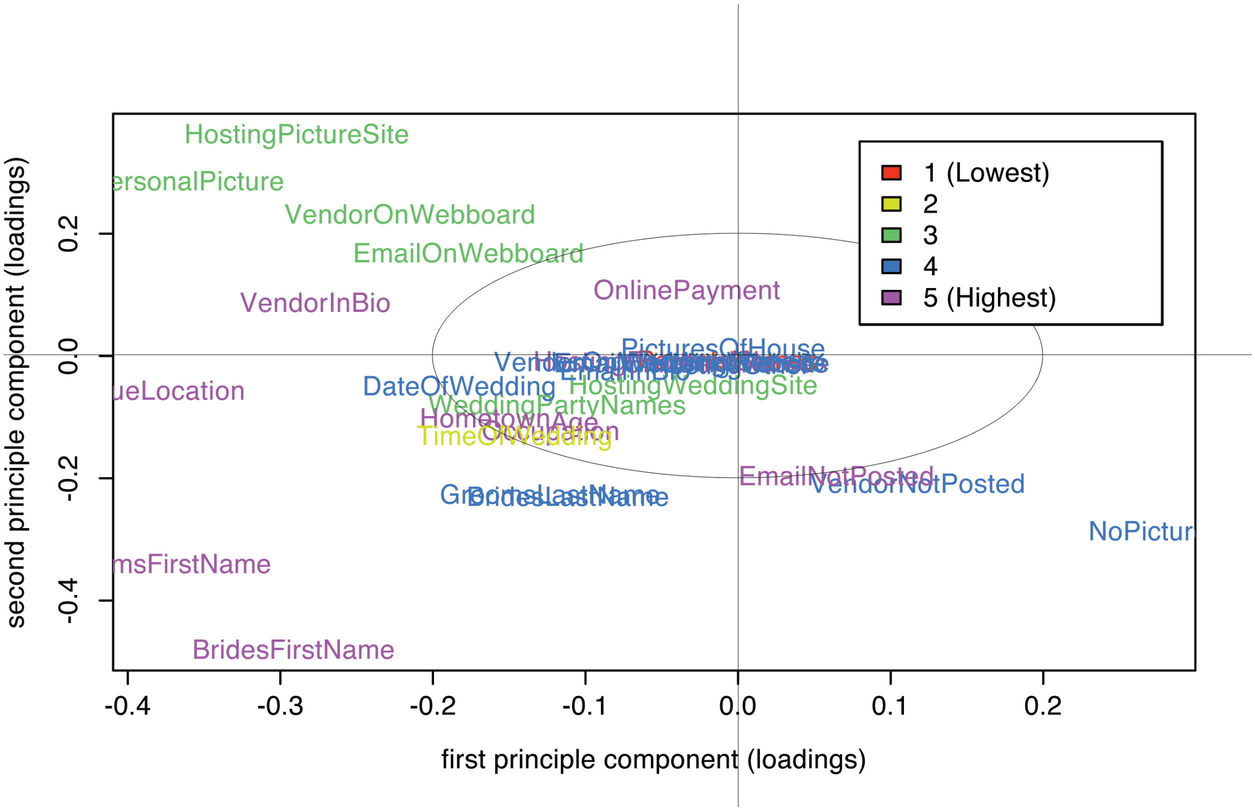


Figure 2: Plot of five frequency ratings on the first two principle components

## 5 Findings

The key findings in our study were:

- There are relationships between how much personal, technological, and wedding information Knot members make available on bios

- A majority of the Knot members surveyed had enough information on their survey to fall victim to at least one privacy concern

- Knot members who did not have any privacy concerns were more likely to have information indirectly related to them, whereas Knot members who were susceptible to privacy concerns had more personal and directly connected information in their bios or posted on a web board

Usually in PCA analysis, the first two principle components are sensitive to overall frequency of the data. However, I did not find this in our plot shown in Figure 2. The variables are colored based on a frequency scale from 1 to 5 as discussed before. If the first principle component (x axis) was sensitive to frequency, then the green colored variables would have loaded lower than the blue variables and the only red variable (lowest rating) would not be in the center of the plot. Similarly, on the second principle component (y axis) the red and yellow variables would not have been plotted in the center.

Instead, I interpreted the first principle component as a measure of whether information was posted in the bio or web board. I see that posting information loads low (left side) and not posting information loads higher. For the second principle component, I categorized the information into personal information, such as names, (loading low) and more technological, such as where pictures were hosted, information loading higher.

In Figure 3, I added lines simulating axis rotation to the loadings plotted on the first two principle components. The plot shows us that there are three axes. Variables plotted along the red axis primarily have to deal with technology, such as where pictures and email was posted. The orange axis has variables that relate to personal information, such as names, occupation, and hometown. The current x axis colored blue, corresponds to wedding information, such as wedding venue name, date of the wedding, etc. These axes show us that if a Knot member posted one variable on the axis, they would be more likely to post other variables on the axis in their bio or on the web boards. For instance, someone who posted the bride's first name would be more likely to post the grooms full name, her hometown, and occupation. I attempted to rotate the axes using factor analysis with oblique rotation, however our matrix is singular and thus cannot be inverted as factor analysis requires.

On the third and fourth principle components we see different patterns of differentiation in Figure 4. The right hand side variables can be categorized as variables that cannot be directly connected with the Knot member. For instance, if I had a vendor information and the hometown, I would not be able to directly connect the information with a particular person. However, if I had a personal picture, age,and occupation - I have information that is more directly connected with

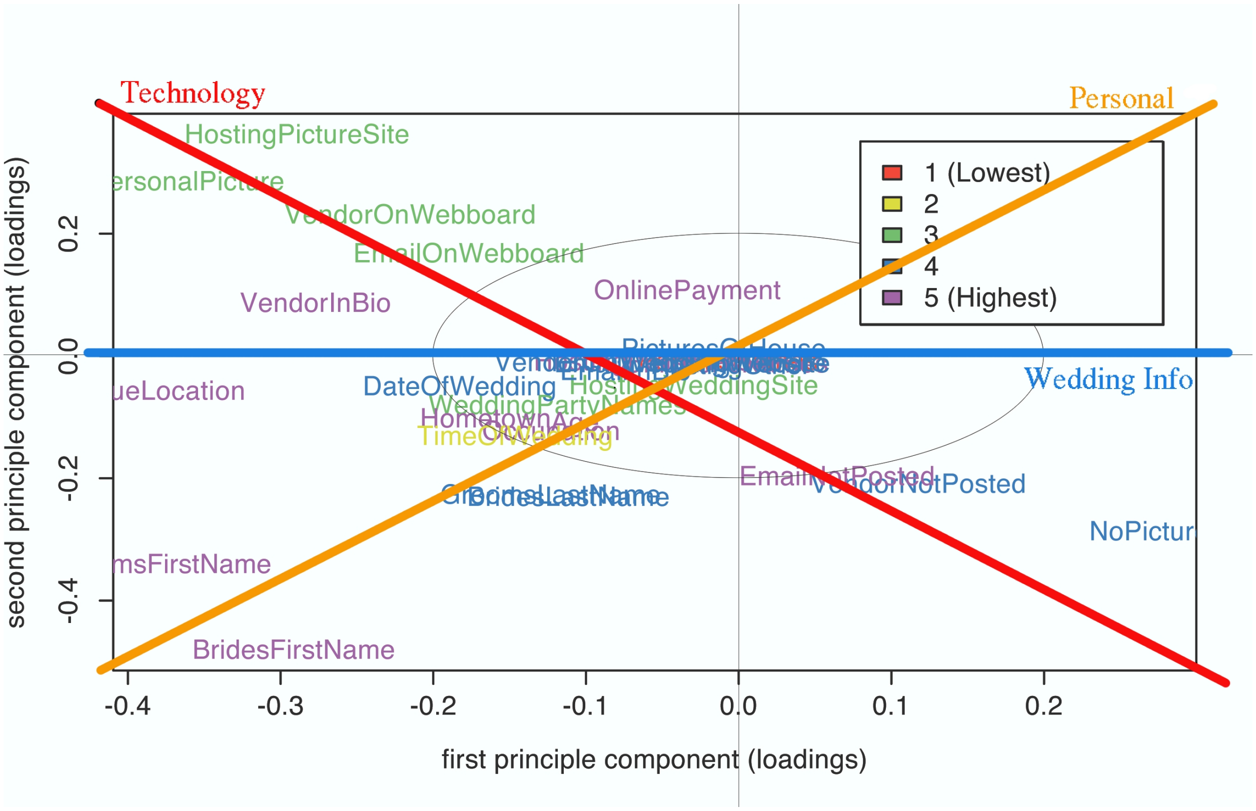


Figure 3: Plot of the PCA loadings with five frequency ratings on the first two principle components with lines simulating axis rotation

an individual. The fourth principle component is the opposite of the second principle component in that I have personal information loading high and technological information loading low.

As discussed in the previous section, I used HCA to create a dendrogram to assist us decide what clusters I should use to color the PCA scores plot. Once I created the dendrogram, I used boolean logic on subtrees to decide how many privacy concerns were on each branch. The dendrogram shown in Appendix B can be broken into three clusters. One cluster primarily had Knot members who were at risk for two or more privacy attacks, another consisted of those who were at risk for one or two privacy attacks, and the third cluster had members who divulged some information, but none of the information had enough to create a privacy attack. The second cluster was not pure because one subtree had members who were not at risk for any privacy concerns ("safe" members).

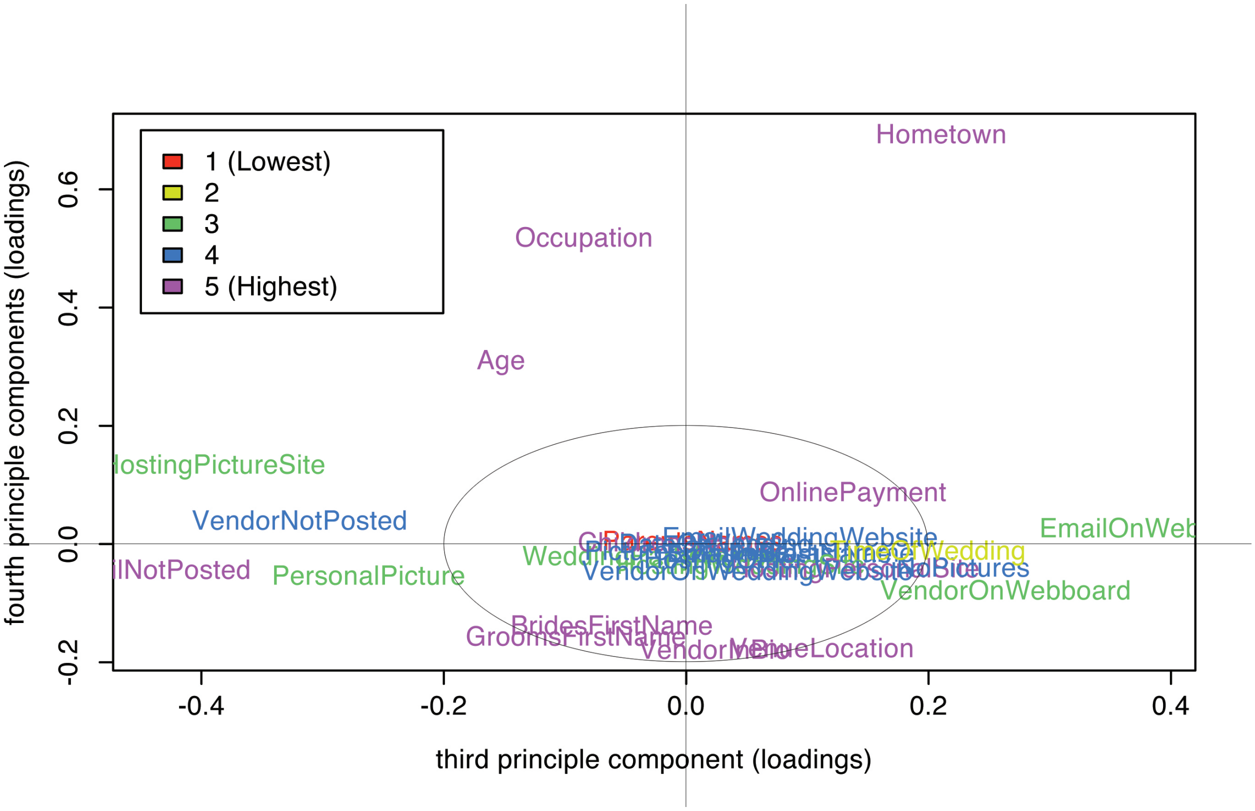


Figure 4: Plot of the PCA loadings with five frequency ratings on the third and fourth principle components

Once I established the clusters from the HCA, I applied the clusters to the scores plot on the first and second principle components as shown in Figure 5. The PCA scores plot clearly shows the first, second, and third cluster and bolsters our interpretation of the first and second principle components. As I can see, the first cluster (the cluster with two or more privacy concerns) has more posted information than the second and third cluster. Some of the second cluster mixes with the third cluster on the right - these Knot members could be the safe members I identified in the second cluster. As shown on the second principle component, all three clusters posted a mix of personal and technological information. However, surprisingly the third cluster that does not post enough information for a privacy concern posts more personal information.

When I plotted the PCA scores on the third and fourth principle components, I see that the first cluster with the most risk of privacy concerns has a mixture of directly and indirectly information

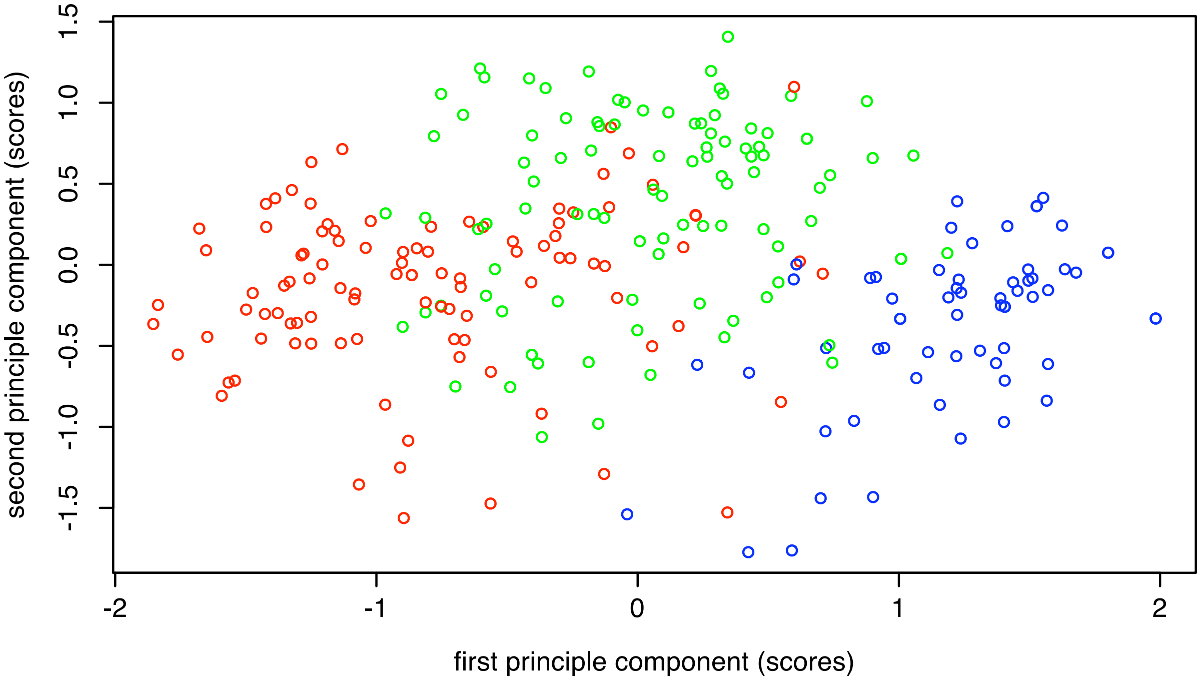


Figure 5: Plot of the PCA scores colored by clusters from HCA on the first and second principle components

and personal and technological information posted. The second cluster has more directly connected information and mostly personal information. Whereas, the third cluster has a mix of personal and technical information, but leans towards indirectly connected information.

## 6 Discussion and Future Work

The plots of the PCA loadings show the relationships between how much personal, technological, and wedding information Knot members make publicly available on their bio or posted on the web board. Knot members are more likely to post first names, venue location, and vendor information in their bios. As suggested by analyzing the first and second components, some members who give their first names will be more likely to give last names, hometown, and the date of their wedding.

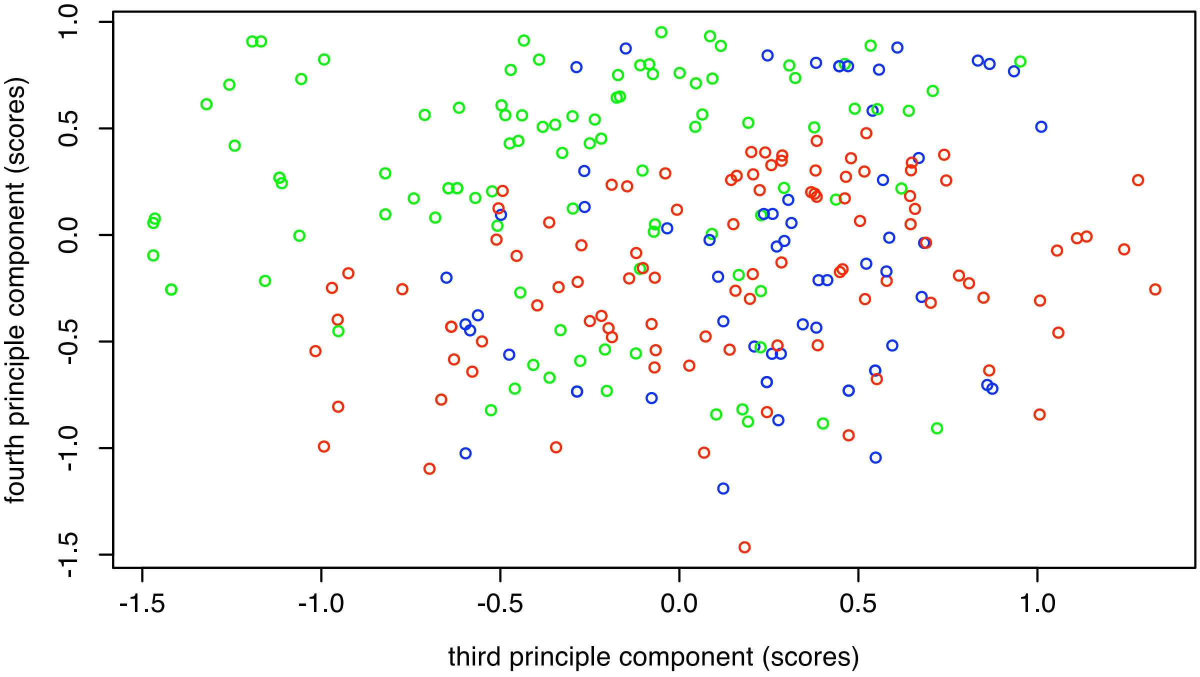


Figure 6: Plot of the PCA scores colored by clusters from HCA on the third and fourth principle components

The PCA scores plot showed us that the more information Knot members post in their bio or web board, the more at risk the member is to have a privacy attack.I found a majority of the Knot members post enough information to have at least one privacy concerns using the minimum amount of data needed for each privacy concern listed in Table 1. However, as Table 2 shows, once I require "helper" information to implement each privacy concern, the percentage of Knot members at risk for attacks decreases drastically.

One disturbing thing that happened while collecting data for this survey is that Knot members participated in the study at all. I simply posted an official looking, HSC approved call for participation and 262 members took the survey. From a privacy point of view - this is unsafe because if I was a malicious third party, I would not have to look through the thousands of bios out there. I could just look for the particular variables I wanted to be true, visit that biography, and then

| Privacy Concern | Min. Info Needed<br>( % at Risk) | Helper Info<br>(% at Risk) |
|---|---|---|
| Phishing | 32.57% | 8.43% |
| Cancellation Problems | 30.27% | 4.60% |
| Identity Theft | 50.96% | 6.90% |
| Robbery | 10.73% | 0% |
| No Privacy Concerns (Safe) | 2.30% | - |

Table 2: Percentage of Knot members surveyed who are at risk of a privacy attack

implement a privacy attack. I am thankful the Knot members participated in my legitimate study, but hope malicious third parties would not take advantage of this generosity and trust.

This privacy study did have some immediate beneficial effects for the Knot community. First, whenever I posted a call for participation, the web board would come alive with messages about how safe people feel about posting information about their weddings online. Second, some members who completed the study wrote in the comments box that after completing the survey, they realized they had too much personal information available. Sometimes when I were analyzing the data and visiting bios with multiple privacy concerns, I realized that the Knot member had already changed their bio to eliminate names, vendors, hometowns, etc.

Even though the survey did have some immediate effects, this is not enough. This study shows that the general public needs to learn more about how to keep their "digital identities" anonymous and protect themselves from privacy concerns. Who is responsible for this education? The Knot? The government? These questions have yet to be answered. However, until these questions are answered, I am willing to edit the current survey to provide feedback to the people who take the survey and advise them on potential privacy concerns. I am also going to contact The Knot with my results and give suggestions on how to educate Knot members.

## 7 Conclusion

Is The Knot a web site that helps couples plan their weddings or a place where members can unexpectedly open themselves up to privacy attacks (the noose)? Our statistical analysis of 262 Knot member bios shows us that some members are susceptible to privacy concerns and that there is a noticeable connection between the types of information divulged in bios and web boards. Education and modified surveys can help protect Knot members from putting too much personally connected information on their bios and web boards.

## References

[1] The knot. http://www.theknot.com.

[2] Kay H. Connelly, Tom Jagatic, Ashraf Khalil, Yong Liu, Katie A. Siek, and Sid Stamm. The internet hunt revisited: Personal information accessible via the web. In *Proceedings of www@10*, 2004.

[3] Federal Trade Commission for the Consumer. Id theft. http://www.consumer.gov/idtheft/.

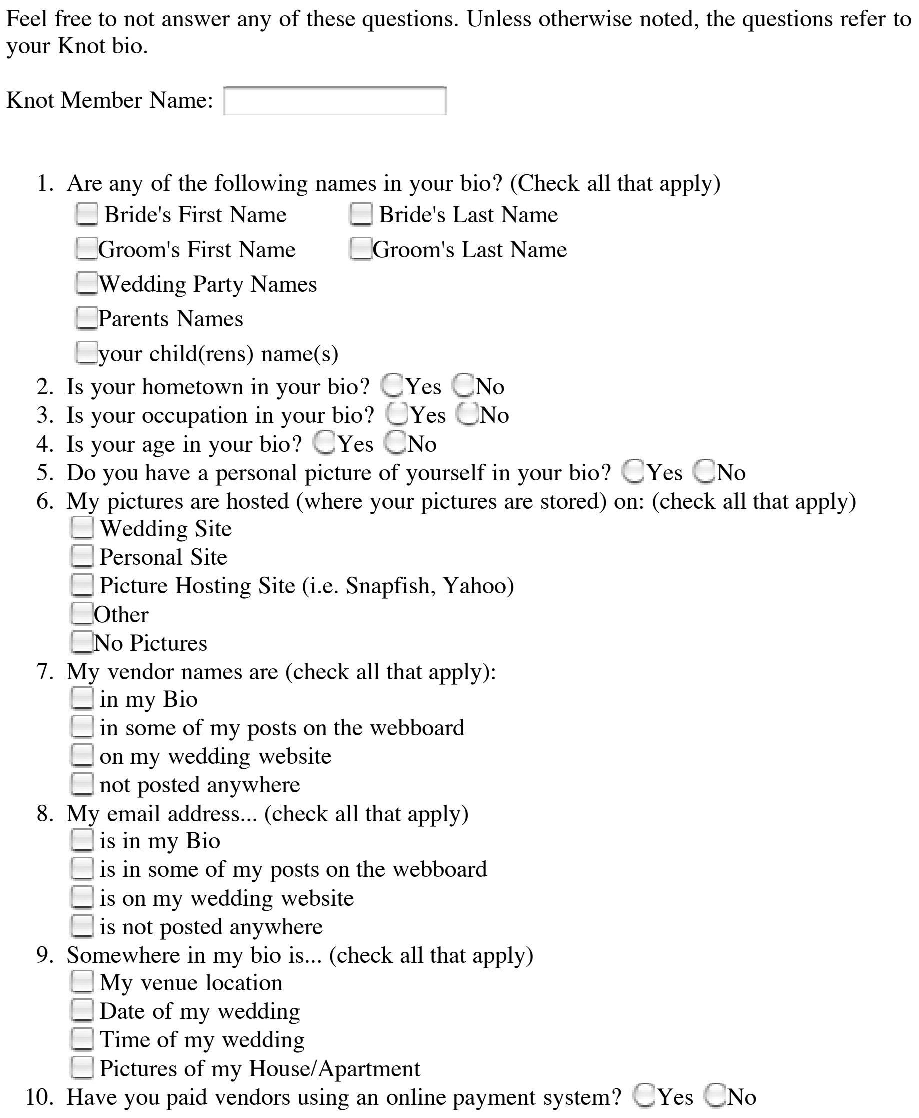
[4] Markus Jakobsson. Modeling and preventing phishing attacks. Phishing Panel of Financial Cryptography '05, 2005.

[5] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-07-0.

[6] United States Secret Service. Public awareness advisory regarding "4-1-9" or "advance fee fraud" schemes. http://www.secretservice.gov/alert419.shtml, 2002.

[7] Lindsay I. Smith. A tutorial on principle component analysis. Web site, February 2002. http://kybele.psych.cornell.edu/l.pdf.

# A Knot Privacy Survey and Data



Feel free to not answer any of these questions. Unless otherwise noted, the questions refer to your Knot bio.

Knot Member Name: □

1. Are any of the following names in your bio? (Check all that apply)
   □ Bride's First Name □ Bride's Last Name
   □ Groom's First Name □ Groom's Last Name
   □ Wedding Party Names
   □ Parents Names
   □ your child(rens) name(s)
2. Is your hometown in your bio? ○ Yes ○ No
3. Is your occupation in your bio? ○ Yes ○ No
4. Is your age in your bio? ○ Yes ○ No
5. Do you have a personal picture of yourself in your bio? ○ Yes ○ No
6. My pictures are hosted (where your pictures are stored) on: (check all that apply)
   □ Wedding Site
   □ Personal Site
   □ Picture Hosting Site (i.e. Snapfish, Yahoo)
   □ Other
   □ No Pictures
7. My vendor names are (check all that apply):
   □ in my Bio
   □ in some of my posts on the webboard
   □ on my wedding website
   □ not posted anywhere
8. My email address... (check all that apply)
   □ is in my Bio
   □ is in some of my posts on the webboard
   □ is on my wedding website
   □ is not posted anywhere
9. Somewhere in my bio is... (check all that apply)
   □ My venue location
   □ Date of my wedding
   □ Time of my wedding
   □ Pictures of my House/Apartment
10. Have you paid vendors using an online payment system? ○ Yes ○ No

Figure 7: Online survey Knot members were asked to complete.

| **Varible from Survey** | **Total Positive Responses** | **Percentage of Positive Responses** |
|---|---|---|
| BridesFirstName | 139 | 53.2567 % |
| BridesLastName | 34 | 13.0268 % |
| GroomsFirstName | 114 | 43.6782 % |
| GroomsLastName | 37 | 14.1762 % |
| WeddingPartyNames | 33 | 12.6437% |
| ParentsNames | 1 | 0.3831 % |
| ChildrensNames | 2 | 0.7663 % |
| Hometown | 131 | 50.1916 % |
| Occupation | 187 | 71.6475 % |
| Age | 202 | 77.3946 % |
| PersonalPicture | 158 | 60.5364 % |
| HostingWeddingSite | 21 | 8.046 % |
| HostingPersonalSite | 27 | 10.3448 % |
| HostingPictureSite | 163 | 62.4521 % |
| HostingOther | 21 | 8.046 % |
| NoPictures | 56 | 21.4559 % |
| VendorInBio | 80 | 30.6513 % |
| VendorOnWebboard | 92 | 35.249 % |
| VendorOnWeddingWebsite | 22 | 8.4291 % |
| VendorNotPosted | 46 | 17.6245 % |
| EmailInBio | 35 | 13.41 % |
| EmailOnWebboard | 83 | 31.8008 % |
| EmailWeddingWebsite | 22 | 8.4291 % |
| EmailNotPosted | 59 | 22.6054 % |
| VenueLocation | 127 | 48.659 % |
| DateOfWedding | 208 | 79.6935 % |
| TimeOfWedding | 38 | 14.5594 % |
| PicturesOfHouse | 6 | 2.2989 % |
| OnlinePayment | 46 | 17.6245 % |

Table 3: All of the Variables from the Knot Survey with the total amount and percentage of positive reponses to the variables.

# B Additional Plots

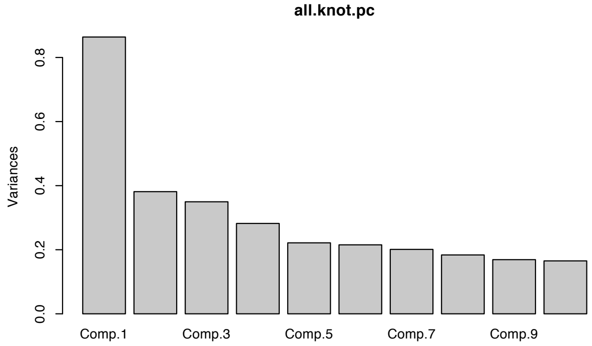


Figure 8: Principle Component Plot

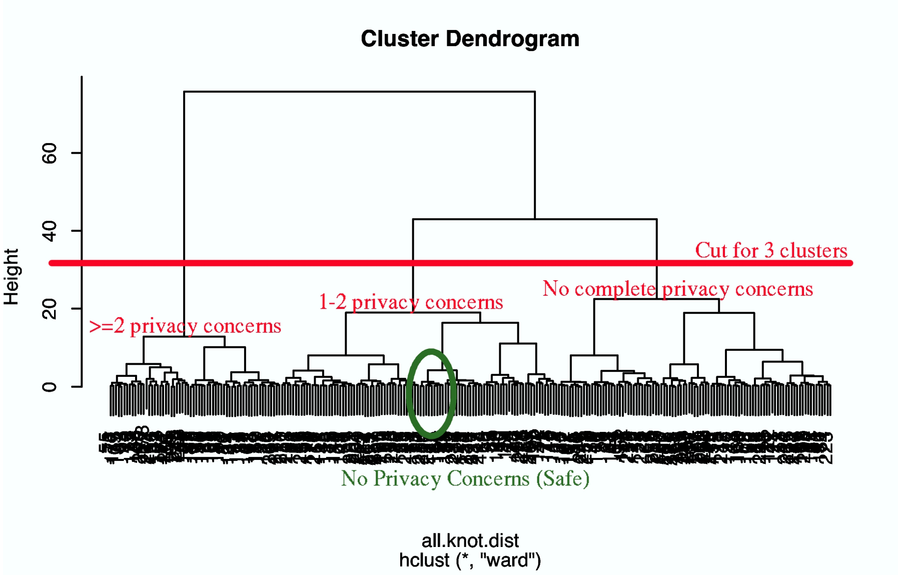


Figure 9: Dendrogram from plotting cluster analysis results