What Blogs Tell Us about Websites: A Demographics Study

Matthew Michelson
Fetch Technologies
841 Apollo St., Ste. 400
El Segundo, CA 90245
mmichelson@fetch.com

Sofus A. Macskassy
Fetch Technologies
841 Apollo St., Ste. 400
El Segundo, CA 90245
sofmac@fetch.com

ABSTRACT

One challenge for content providers on the Web is determining who consumes their content. For instance, online newspapers want to know who is reading their articles. Previous approaches have tried to determine such audience demographics by placing cookies on users' systems, or by directly asking consumers (e.g., through surveys). The first approach may make users uncomfortable, and the second is not scalable. In this paper we focus on determining the demographics of a Website's audience by analyzing the blogs that link to the Website. We analyze both the text of the blogs and the network connectivity of the blog network to determine demographics such as whether a person "is married" or "has pets." Presumably bloggers linking to sites also consume the content of those sites. Therefore, the discovered demographics for the bloggers can be used to represent a proxy set of demographics for a subset of the Website's consumers. We demonstrate that in many cases we can infer sub-audiences for a site from these demographics. Further, this feasibility demonstrates that very specific demographics for sites can be generated as we improve the methods for determining them (e.g., finding people who play video games). In our study we analyze blogs collected from more than 590,000 bloggers collected over a six month period that link to more than 488,000 distinct, external websites.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*clustering*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis*

General Terms

Algorithms, Experimentation

1. INTRODUCTION

A Website's ability to infer information about its audience can directly impact how well that Website caters to its audience (or specific sub-audiences). For instance, online newspapers can provide more interesting articles and more relevant advertisements if they know some of the demographics of their readers (e.g., most of their readers are married women who like cars). Therefore, a problem of interest to many content providers on the Web is discerning some of the demographics about their users.

We define a demographic as a descriptive, binary-attribute about a user. Examples might be whether a user "is-married," or "has-pets." For continuous or real values, a demographic is created by binning the data into sets. For instance, users' ages can become the demographics, "Age 0-10," "Age 11-20," etc. Demographics define subsets of users based upon which of the attributes they share. For instance, the set of users that "has-kids" and "is-male" defines the subset that we generally classify as "fathers." For a Website, then, the goal is to discover different subsets of users defined by different demographics.

Specifically, determining users' demographics can better align consumers and producers of content. Not only does this make the content more targeted for the consumers, it can also lead to improved revenue for content producers. This has spurred a long history into this topic (in the physical world), which is sometimes referred to market research. On the Web, two common approaches for this market research are to leverage users' cookies or query logs or to directly ask consumers about themselves (e.g., via some survey). Each of these have certain limitations and drawbacks which we seek to overcome in this paper.

Commercially, many companies place cookies on users' computers that track statistics such as which sites they visit, etc. Such data can then be mined using techniques such as Web usage mining (e.g., [13]) to try and discern what users might find interesting, based upon their browsing history. For instance, such methods can provide a personalization service, recommending new content based upon what a user seemed to like previously.

Such techniques define pseudo-demographics in that they can tell users that "people like you also like this Website," but they are not true demographics in that attribute based profiles of a user cannot be generated. Rather, these mining techniques exploit latent demographics in that they are exploiting some demographics about the user, based upon his or her past behavior. However, the end result is not a clearly defined set of demographics about the user. Therefore, these methods cannot be used by content providers to understand attributes about their audience. While this is good to maintain the status quo, it makes it harder for a content provider

WSDM'11, February 9–12, 2011, Hong Kong, China.

to broaden their appeal, or build stronger loyalty by addressing other topics that might appeal to the demographic of its audience. For example, the latent behavior might suggest that the audience likes cars of a certain type. However, knowing that this audience consists of males who are likely married with kids may also suggest other types of content that would be welcomed. In other words, methods using cookies might make it hard to target good content for well-defined users because they cannot really tell who the users are.

Further, and equally important, many users are wary of cookies or any persistent identifiers used to track them [1]. This is largely due to lack of transparency and privacy-concerns as cookies have been noted to lack "informed consent." That is, users are sometimes unaware how and when cookies are being used, and they lack easy options to block them [10]. Therefore, using cookies to gain demographic insight about users can actually end up diverting the consumers and producers of content.

The other popular option to gain demographic insight is to leverage surveys. However, this may not be scalable. First, previous work explains that users generally do not like to share personal information with Websites [6, 2]. Second, because of the low response rate for surveys, they may require too many users to develop niche demographic profiles or to be used for very large scale studies of an audience.

Instead, our approach leverages a source of data that is both large, to overcome the scalability issue with surveys, and publicly available and open, to alleviate some of the users' concerns about transparency with cookies. Specifically, building upon recent methods for mining blog posts [14], our method determines the demographics of websites by analyzing the posts of bloggers that link to them. In our approach we analyze both the content of the blogs and their network connectivity to determine demographic information about the audience of websites by inferring demographic information about the bloggers linking to the sites. Presumably bloggers linking to sites also consume the content of those sites, and therefore they represent an easily identifiable proxy for some subset of a Website's consumers.

We acknowledge that our approach will likely identify a subset of a Website's consumers, which might not necessarily represent the entire audience breakdown for a Website. However, bloggers generally have an active online presence and a willingness to voice themselves publically (both by definition), and therefore they represent an influential subset of a Website's audience. We feel such representative groups represent an important segment to identify as they will be more likely to drive traffic to a site via their blog.

One of the important aspects of our approach is that we are demographically agnostic. We first build up demographic profiles of each blogger, based upon analyzing the text of the blogger's posts, and then analyze the sites they link to, to identify the demographics for those sites. In this manner, as new or more specific demographics are discovered via better text analysis, those can be immediately applied to discover demographics of sites. That is, as we develop text analyzers to discover niche demographic attributes, such as whether someone works in A.I., we can immediately apply these demographics to the Websites.

We make three main contributions in this paper.

1. We define an approach for determining the demographics of Websites by analyzing the blogs that link to them. We combine text analysis methods with social network analysis methods and develop a descriptive set of demographic attributes that apply to the Websites.
2. We define a flexible approach that is not tied to specific demographics, but that can easily incorporate new demographics, as long as the demographics can be discovered by analyzing blog text.
3. We present a large scale study of this problem using a corpus of data collected over a six-month period representing over 590,000 bloggers.

The rest of this paper is organized as follows. Section 2 details our approach for leveraging blogs to determine Website demographics. Section 3 describes our experimental study. Section 4 provides related research, and Section 5 presents our conclusions and future work.

2. SITE DEMOGRAPHICS FROM BLOGS

In this section we describe our approach to discovering the demographics for a subset of a Website's users by analyzing the blogs that link to the Website. More specifically, we analyze both the text of the blogs and the network connectivity with the aim of inferring some demographics about the users linking to sites, under the assumption that some of these demographics will apply to the linked site as well. Our whole approach, called "Blographics," focuses on identifying the demographics for a particular site and breaks into three high-level steps as shown in Figure 1.

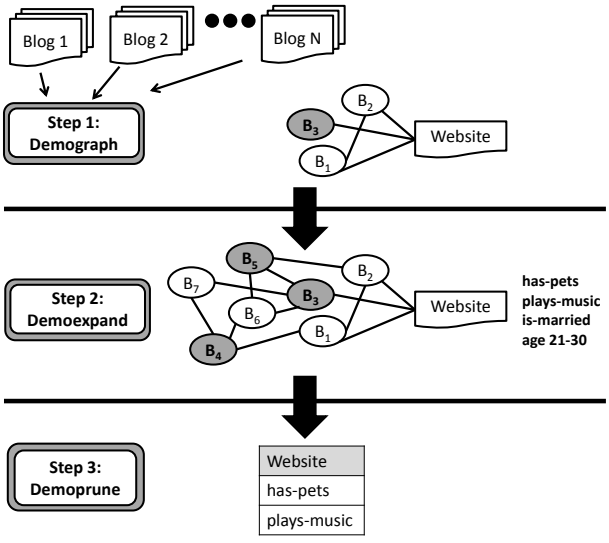

Figure 1: Blographics: Our overall approach. Shaded vertices denote bloggers where some demographic information was identifiable.

Let us consider the complete blogosphere as the graph $\mathcal{G} = (V, E)$, where V is the set of all bloggers and the Websites they link to and $E = \{e(i,j)|v_i, v_j \in V\}$ is the set of all edges between vertices in V. Further, each vertex $v_i \in V$ contains demographic information represented as a k-ary vector representing k demographic attributes. This vector is represented as $\mathbf{x}_i = \{x_{i,1}, \ldots, x_{i,k}\}$, where $x_{i,j} \in \mathcal{X}_j$, is the value of demographic attribute i, which can take on a

value in $\mathcal{X}_i$. For example, a demographic attribute can be "is-married" and the value is binary. Another attribute may be the age of the blogger, etc. We represent all vectors of V_w as $\mathbf{X}_w$. For blogger vertices, the demographic information is specific to that blogger and is often not directly observable. In fact, the demographic information is often not at all observable or inferable from a blogger's profile or the posts of the blogger. For Website vertices, the demographic information is an aggregate of the demographics of the bloggers linking to it. Our "Blographics" approach is focused specifically on *deriving* this aggregate demographic vector for a given site w.

In the first step, called "Demograph" we create a social-network graph around a particular Website w which consists of the bloggers directly linking to w and their demographics. We refer to this graph as $\mathcal{G}_w \subseteq \mathcal{G}$, where $\mathcal{G}_w = (V_w, E_w)$. As before, V_w are the vertices in $\mathcal{G}_w$, which represent the set of bloggers directly linking to w, and $E_w = \{e(i,j)|i,j \in V_w\}$. As such, this first step breaks into two parallel steps: identifying $\mathbf{x}_i \in X_w$, the demographics for each individual blogger, and finding E_w, the connections between bloggers who link to site w.

The size of V_w is often small and the number of bloggers with identifiable demographic information is even smaller. We therefore expand on this set to pull in more related bloggers. Specifically, we leverage a commonly known characteristic of linking behavior between people called *homophily* [9]: people who share characteristics are more likely to connect than those who do not. The more you have in common the more likely you are to link with each other. We leverage this phenomenon in this step, called "Demoexpand," which expands the set of users V_w to include their immediate neighbors and their links. We denote this new graph as $\mathcal{G}_w^1 = (V_w^1, E_w^1)$, where the 1 superscript means 1 hop away from w, which can be generalized to any $n \geq 0$ where $\mathcal{G}_w^0 = \mathcal{G}_w$. This larger graph often includes many more bloggers, some of which have identifiable demographic information. Due to homophily and statistics, the aggregate demographics of this larger set of bloggers represent a good proxy for the demographics of site w. Clearly, this larger set is not as perfect a match as the directly linking bloggers, and the further away one travels from these directly linked bloggers, the less demographic agreement we would expect with the original set. We therefore limit ourselves to only the 1-hop neighbors.

At the end of the Demoexpand step we have a group of bloggers who link (in)directly to w, giving us possible demographics for that site. In order to clearly identify the demographics for this particular site, however, we must differentiate them from a baseline representing bloggers in general. It may be that many bloggers linking to a site are male and married, but if that is true in the blogosphere in general, then this is not a particularly useful demographic to identify for this site. Rather, we are interested in identifying demographic information which is specific to this site and different from the blogosphere at large. Therefore, the last step, "Demoprune," eliminates demographics for sites that are not applicable. The end result is a set of demographic attributes for Websites, based upon the bloggers that link to them. Figure 1 shows our overall workflow. We now detail each of these steps.

2.1 Demograph: Discovering demographics

The first step in our approach, Demograph, takes as input a large set of blogs and a Website w. Demograph outputs $\mathcal{G}_w$, a graph representing the bloggers directly linking to w, along with their demographics. We break this down into two substeps, which we accomplish in parallel: discovering the demographics for each blogger, and defining the initial social graph of bloggers linking to Website w.

As described above, the graph's vertices are either individual bloggers or Websites, and each blogger vertex may have some demographics assigned to it. The edges represent direct links between bloggers (forming a social network) and links between a blogger and Website w as when the blogger links to the site. To create the graph, we analyze the set of blog pages for each individual blogger. We generate all outgoing edges for a blog vertex by extracting all of the links from the individual blogger's pages, keeping only links to other bloggers or to Website w. We prune these links to only include links to other blogger also linking to w.[1]

In parallel, we discover the demographics for each blogger. This generates $\mathbf{x}_i$, the demographic attribute vector of the blogger vertex v_i which will contain the demographic information we can discover about blogger i from the blog pages themselves. Specifically, following prior work [14], we generate demographic facts for each blogger by analyzing the text of their blog. Our goal is to discover precise facts about users that clearly map to demographics. This precision is important to limit noise, and we here leverage the fact that we can monitor many bloggers thereby generating enough identifiable demographic information in the aggregate. One set of facts that can be extracted with high precision and that map easily to demographics are "self-referring" facts. A self-referring fact is a sentence that a person writes about themselves that very, clearly delineates a fact about themselves. For instance, if a blogger writes, "My wife came home from work at 5," then we can infer that the blogger is married, and therefore has the demographic "is-married." Self-referring facts are largely unambiguous (as we define them) and lead to a clean, but small, set of demographics for each user. Therefore, we define a set of self-referring facts by mapping the facts to specific demographics, and then perform extraction from the text using patterns that represent the facts. For instance, to discover which bloggers have the "is-married" demographic attribute, we look for phrases such as "My wife," "My spouse," etc., which directly map to the demographic attribute. The constraint is that we need strict methods for finding demographic attributes because we want to ensure their validity.

As we note above, our approach is agnostic to the specifics of the actual demographic attributes. Therefore as more sophisticated methods for identifying demographics from text become available they can easily fit within this methodology allowing us to create more detailed and fine-grained sets of demographics to assign to the bloggers (e.g, previous work demonstrates that text analysis can identify bloggers' native language [7], age and gender [12]).

To summarize, the complete Demograph step analyzes the set of blog pages for each blogger, creating a vertex in a Website-specific social graph which represents the blogger

[1]We note that we do generate the full blogosphere graph ($\mathcal{G}$) for later steps and analysis, but we consider only $\mathcal{G}_w$ for this part of our approach.

and the links from the blogger to Website w and to other bloggers linking to w. Further, each blogger vertex (possibly) contains demographic information discovered by extracting self-referring facts. Figure 2 demonstrates this process graphically.

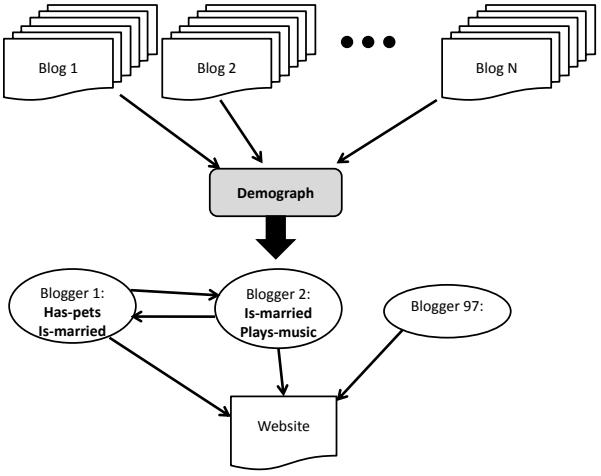

Figure 2: Demograph

2.2 Demoexpand: Using network analysis to improve demographic fidelity

The previous step builds $\mathcal{G}_w = (V_w, E_w)$, a social graph of bloggers directly linking to Website w. As noted above, a small subsample of V_w will contain identifiable demographics information from the blog posts of bloggers in V_w. However, the number of bloggers linking to w (i.e., $|V_w|$) is often quite small and follows an extremely long tail distribution (as we describe in our experiments). Therefore, we need to increase the size of V_w–the number of bloggers relevant to Website w. This will help alleviate some of the sparsity, allowing us to infer more demographic attributes for w. To do this, we rely on "homophily" (cf. [9]) to expand the set of users that can help us infer demographic information about w. Specifically, we expect that through homophily bloggers that link to each other likely share similarity and therefore represent an expanded set of consumers for that particular site.

Therefore, we here leverage homophily by expanding $\mathcal{G}_w$ to include not only the set of bloggers that link to site w but also any blogger which may link to (or be linked to by) $v_i \in V_w$. We call this "link expansion." That is, if we consider the complete blogosphere $\mathcal{G} = (V, E)$, expand V_w to include all vertices in V_w and their neighbors. More formally, define N_i as the neighborhood of v_i which consists of all bloggers linking to v_i as well as all bloggers linked to by v_i. We then define $V_w^1 = \bigcup_{v_i \in V_w} N_i$. By definition, $V_w \subseteq V_w^1$. We note, this step can be generalized to identify any $V_w^n = \bigcup_{v_i \in V_w^{(n-1)}} N_i$, $E_w^n = \{e(i,j) | (v_i, v_j) \in V_w^n\}$, and $\mathcal{G}_w^n = (V_w^n, E_w^n)$. However, we limit ourselves to consider only $\mathcal{G}_w^1$ as moving further away from w will pull in bloggers who have a decreasing interest in w. The link expansion step is shown graphically in Figure 3.

By expanding the set of bloggers that link to site w, we created a larger pool of bloggers from which to infer demographics for the site. Therefore, the final step is to select the demographics, based upon the bloggers, that are appropriate for w. In turn, these demographics will become the descriptive and differentiating demographics for w.

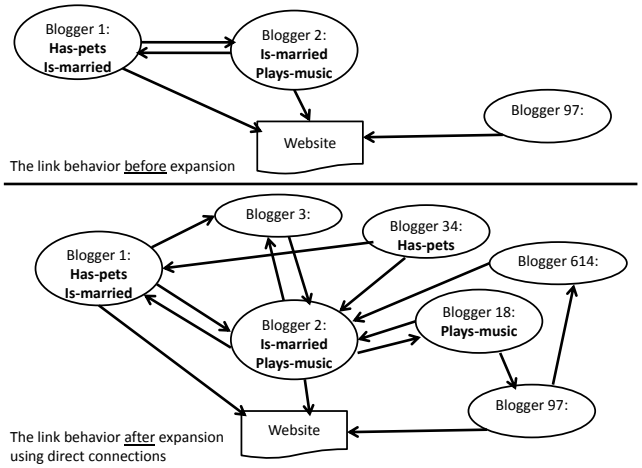

Figure 3: Link expansion to Websites

2.3 Demoprune: Finding relevant demographics for a Website

The output of the previous steps is $\mathcal{G}_w^1$, which contains a set of bloggers that represent some subset of potential consumers for a Website w. V_w^1 denotes the set of bloggers in $\mathcal{G}_w^1$ and $\mathbf{X}_w^1$ is the set of all demographic information identified from $v_i \in V_w^1$. The goal of this step is to discover the demographics of that subset that apply specifically to w–i.e., the demographics which are specific to w and not to the blogosphere at large. We previously defined a demographic attribute as a particular attribute that applies to a person (or blogger). We now define a "specific demographic" as a demographic attribute that applies more significantly to the set of bloggers in V_w^1 than it does to bloggers in general. *Specific demographics* are defined by their proportion in a sub-community versus their proportion in the community as a whole. For instance, amongst a set of blogging computer scientists, a demographic such as "Publishes AI Papers," would be a specific demographic because it would apply with a much higher proportion amongst the group of blogging computer scientists than it would to bloggers in general.

Our approach to discovering specific demographics for w is therefore to look for demographics that occur more frequently in V_w^1 than for bloggers in general. Essentially, this is the problem of discovering population proportions that are significantly different for the subset than from all bloggers. In our method, we discover specific demographics for w employing the following constraints:

1. The proportion of the demographic attribute being observed in V_w^1 is larger than the proportion of it being observed in the blogger population in general.

2. The increased proportion represents a statistically significant difference, using the z-test for population proportions.[2]

[2]We use 95% confidence levels for our significance testing.

We identify the specific demographics by first defining demographics and their proportions for our whole set of bloggers (which represents a random distribution of bloggers in general). Then we compare the demographics and their proportions within V_w^1, and keep only those that fit our pruning constraints, assigning them as demographics for w.

The algorithm is given in Table 1. The inputs to the algorithm are $\mathcal{X}$, the set of Demographic Attributes; w, the Website of interest; B_w the bloggers linking to w (defined as either V_w^1 or V_w, depending on the specifics of the experiment as we describe below); and B, the complete set of bloggers in the blogosphere. First, we compute the baseline proportions for each demographic using the whole blogger population. Then, we compare the demographic as it applies to the bloggers that link to w, versus the baseline bloggers. If the proportion represents a statistically significant increase, then we assign that demographic to w as a specific demographic.

Table 1: Finding specific demographics

DEMOPRUNE
Input: $\mathcal{X}$, w, B_w, B

$X_w \leftarrow \{\}$
For each $X_i \in \mathcal{X}$
  P_w =PROPORTION(X_i,B_w)
  P_B =PROPORTION(X_i,B)
  if ($P_w > P_B \wedge$ Z-test(P_w, P_B))
    $X_w \leftarrow X_w \bigcup X_i$
Return X_w

3. EXPERIMENTAL STUDY

In this section we validate that we can leverage blog demographics to determine Website demographics.

Our experimental data consists of blogs collected from 590,011 bloggers across a six month period from 1-29-2010 to 7-13-2010. This set of bloggers link to more than 488,000 distinct, external Websites. Because our data set is limited to just six months, we noticed a significant long-tail distribution in terms of the number of bloggers linking to sites, with almost all sites having just one or two bloggers linking to them. Therefore, to generate more meaningful demographics, for this study we analyze only those sites that are directly linked to by at least 10 bloggers. This reduced the number of sites under consideration to 21,121. However, the sample size of bloggers linking to these sites is much more meaningful.

Our test demographics represent both standard and niche demographics, and all are defined via self-referring facts. Table 2 lists the test demographics we look for, each of which has a name and some example phrases we used to identify them from text. Again, the goal is discover highly precise demographics, but their restrictive nature implies that their coverage may be small (e.g., the identification of the demographics has very high precision at the expense of recall). We note that we aim to discover both standard demographics (such as sex or age) and niche demograhpics which may be important for cases where one wants to identify very specific subsets of users (such as those who "own pets" or are "grandparents"). Both are important in different domains and applications where either generality or specificity may be more important (e.g., one advertisement may focus on a general audience, such as "older men" but another may target "pet owners who play music.") The key insight for this work is that we are essentially agnostic as to the types of discoverable demographics, such that arbitrary demographics can be defined, discovered, and used. This allows our method to be flexible regarding the required demographics.

Table 2: Demographics and self-referring facts

Demographic Name	Example phrases
In relationship, male	"My girlfriend ...," "My wife ...,"
In relationship, female	"My boyfriend ...," "My husband ...,"
Age 0-10, 11-20, ..., 60+	"My 21st birthday ...,"
Is Grandparent	"My grandson ...,"
Is Parent	"My kids ...,"
Is Married	"My spouse ...,"
Plays music	"My guitar ...,"
Has Pets	"My dog ...,"
Mentions parents	"My mum ...,"
Mentions grandparents	"My grandpa ...,"

We calculated the proportions of these demographics for our baseline population of all 590,011 bloggers in our data set. There are two special cases to note. The proportions for "In relationship, male," and "In relationship, female" are calculated based on the identified sample. That is, if #M is the number bloggers we identified as men in a relationship, and #F is the number of bloggers we identified as women in a relationship, we define the proportion for "In relationship, male," as #M/(#M + #F). Then, we include another demographic, "Unknown Gender," to represent the rest of the population that we could not identify as fitting into either of these groups. That is, if our total population is N, then we define "Unknown Gender," as N-(#M+#F)/N. We also perform the same calculation for each age range. The rest of the attributes are binary (either they occur or do not) and we define their proportions as bloggers that have the attribute over the size of the population. We treat these differently because people rarely mention that they do not have a particular attribute, and we therefore cannot have an unknown for these types of attributes–in other words, we only have positives. However, if the attribute is mentioned, then clearly this is significant for the blogger mentioning it and it is therefore of particular interest for identifying specific demographics.

Table 3 shows each demographic, the percentage of our total population of bloggers with each demographic, and the absolute number of bloggers that the percentage represents. This serves as a baseline for determining the specific demographics for each subset of bloggers that link to a site. One thing to note is that even though some of these attribute percentages are small, we still find that they apply to certain Websites. This indicates that a site strongly identifies with bloggers who have that attribute, as it applies to significantly more bloggers linking to the site than the population as a whole. That is, although the results may seem sparse, there is enough signal from the demographic identification to be used to determine Website audiences.

Table 3: Baseline demographics

Demographic	Proportion (%)	Num. Bloggers
In relationship, male	42.18	10,004
In relationship, female	57.81	13,712
Unknown gender	95.98	
Age 0-10	9.12	54
Age 11-20	17.89	106
Age 21-30	41.75	246
Age 31-40	13.33	79
Age 41-50	7.72	46
Age 51-60	3.86	23
Age 61-70	2.81	17
Age Unknown	99.90	
Is Grandparent	0.19	1,121
Is Parent	8.86	52,275
Is Married	4.28	25,252
Plays Music	0.57	3,363
Has Pets	4.04	23,836
Mentions Parents	18.68	110,214
Mentions Grandparents	2.12	12,508

3.1 Demographics of $\mathcal{G}_w$ vs $\mathcal{G}_w^1$

Our first experiment studies our ability to generate specific demographics for a Website, based upon the bloggers that link to it. Again, these specific demographics are those that are statistically significant improvements over the percentages for those demographics in the baseline population.

Our hypothesis is twofold: first, that we will get more identifiable demographic information about bloggers by expanding from $\mathcal{G}_w$ to $\mathcal{G}_w^1$, and second, that these extra demographics will yield cleaner and more significant results. To gauge the first part of our hypothesis, we first investigate how many bloggers are pulled in when going from $\mathcal{G}_w$ to $\mathcal{G}_w^1$. Figure 4 shows the increase in bloggers across all the 21,000 Websites in our study. As we can see, around 12,000 sites had an increase on the order of 100-1000 users, although thousands of Websites also had increases on the order of 1000's and even 10,000's. This validates our hypothesis that $\mathcal{G}_w^1$ indeed includes a much larger set of bloggers.

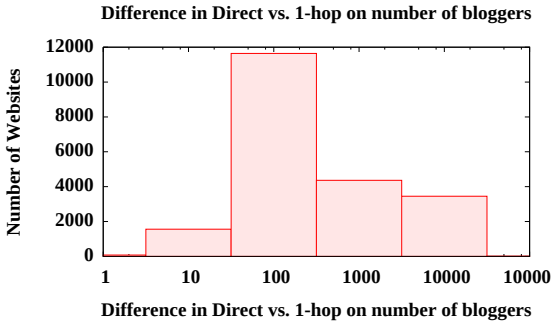

Figure 4: The change in number of bloggers as we move from $\mathcal{G}_w$ to $\mathcal{G}_w^1$. As we can see, 12,000 of the 21,000 Websites had an increase on the order of 100's.

Second, we investigate whether this increase in bloggers also means that the bloggers are more connected. Higher connectivity suggests that there is a tighter blogger community and hence "homophily" may in fact help us. To test this, we investigate whether the number of connected components (considering only edges between bloggers in V_m^1) decreases or stays the same. If the number decreases, then fewer, but more connected, sub-communities are forming. Figure 5 shows how the number of components change as we move from $\mathcal{G}_w$ to $\mathcal{G}_w^1$. As the figure shows, roughly 4,500 Websites (21%) had no change in components. However, the remainder of the Websites (79%) had a decrease in components. Therefore, these Websites are linked to by tighter communities. The tail is actually slightly longer than shown, as the largest decrease in components was 181. However, we removed the longer tail for readability.

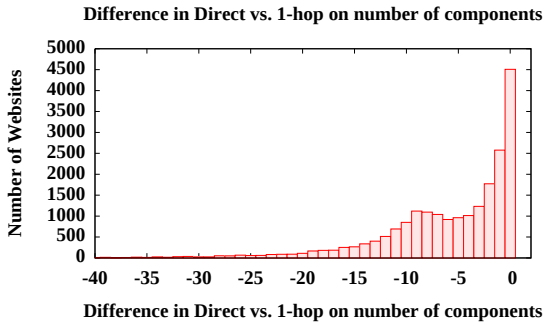

Figure 5: The change in number of connected components among the bloggers as we go from $\mathcal{G}_w$ to $\mathcal{G}_w^1$. Roughly 21% had no decrease, whereas the remainder had decreases in the number of components, showing a tighter connectivity as we pull in more of the community. We cut off the graph at -40, and do not show the long tail terminating at the largest decrease of 181 components. This was done to make the graph more readable.

To add clarity to this phenomena, as we shift from $\mathcal{G}_w$ to $\mathcal{G}_w^1$, we gain new bloggers (nodes) which connect to bloggers (nodes) in one or more existing components. These newly added bloggers may act as bridges between components, connecting them, and resulting in a more larger component that subsumes the previously independent components. Figure 6 shows this graphically. In the top of the figure we see an example site, along with two independent components. In the bottom of the figure, as we add in new bloggers from $\mathcal{G}_w^1$ we see that the newly added blogger forms connections to nodes in previously independent components, resulting in a single, subsuming component.

To test the second part of our hypothesis, we implemented two versions of Blographics which we use to study the implications of both using text analysis to determine demographics, and then extending this approach using the social network analysis. Our first version, called "Direct," generates demographics solely by analyzing the text of the bloggers who link directly to Website w (i.e., bloggers in $\mathcal{G}_w$). This method represents the ability to determine demographics just based on analyzing the text of the bloggers that link to a non-blog website w. The second version, called "One Hop," uses link expansion (Demoexpand) to create $\mathcal{G}_w^1$ and we use this graph to identify specific demographics. In other words, the "One Hop" approach generates demographics by

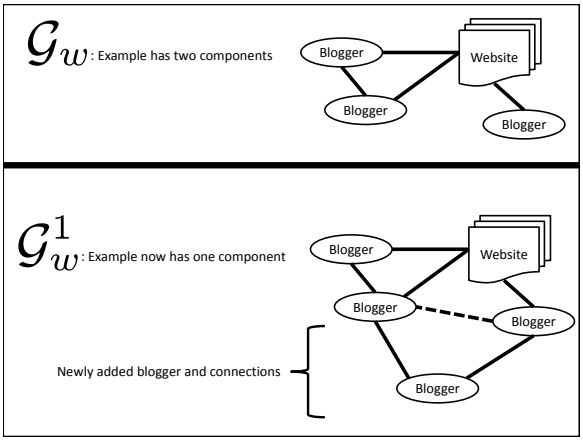

Figure 6: Going from $\mathcal{G}_w$ to $\mathcal{G}_w^1$ yields fewer components as previously independent components become connected.

analyzing the text of the bloggers who link directly to the sites along with their immediate neighborhood. This approach allows us to analyze the improvement by considering the network analysis as well.

We validated our hypothesis and approach in two ways with this experiment. First, we generated all of the specific demographics using the Direct method, which tests the ability to discern demographics based on text analysis. Second, we ran the One Hop approach, and recorded the cases where the new method gained demographics over the Direct method. This demonstrates the improvements we can get using the extra network analysis.

In total, across both approaches, we were able to identify specific demographics for 15,780 of our study Websites, meaning we could generate demographics for 74.71% of them. Table 4 breaks down our results in detail. The first row of the table shows the number of Websites covered by the Direct method, along with the number of discovered demographics, and the average number of demographics discovered per Website. The second row shows the number of new Websites additionally covered by One Hop, along with the number of newly generated demographics and their average over the new sites. The Direct method presents a compelling case for determining demographics using the text analysis of blogs. In partiuclar, a majority of the sites were covered by this method alone, and it generates a substantial number of demographics per site that it covers. The One Hop method adds coverage to almost 50% more Websites, adding in a significant number of demographics as well. As it averages more demographics per site it is able to provide more demographic information. Therefore, indeed we can provide much broader coverage and generate more demographics by leveraging the social network.

In fact, the One Hop method actually finds many types of different demographics that are not as well represented as discovered by the Direct method. Table 5 shows the distributions for the types of specific demographics discovered by each method. The Direct method represents ages and gender (based on relationships) more, emphasizing those types. The One Hop method seems to discover more people with pets and more married people, proportionally. In fact, the KL-Divergences between the Direct and One Hop method is 1.71, showing that indeed the distributions differ and emphasize different demographic types. Therefore, the One Hop method is useful for not only discovering more demographics, but also different types of demographics.

Table 4: Discovered demographics ("One Hop" lists the number of demographics found in *addition* to the one's found by the "Direct" method.)

Method	Total Sites	Total Demographic Attributes	Average Demographic Attributes per Site
Direct	11,500	26,025	2.26
One Hop	4,280	14,061	3.29

Table 5: Discovered demographic distributions

Demographic	Direct	One Hop
In relationship, male	4.23	0.03
In relationship, female	20.08	0.75
Age 0-10	0.56	0.23
Age 11-20	0.76	0.26
Age 21-30	0.35	0.00
Age 31-40	0.86	0.68
Age 41-50	0.51	0.54
Age 51-60	0.19	0.18
Age 61-70	0.09	0.14
Is Grandparent	0.18	0.26
Is Parent	17.28	18.63
Has Pets	7.42	22.15
Is Married	7.78	20.72
Plays Music	0.80	7.82
Mentions Parents	35.18	10.95
Mentions Grandparents	3.72	16.67

3.2 Example Website Demographics Found

The previous results demonstrate that we can find specific demographics and that by leveraging network analysis, we can improve this process. Therefore, as text analysis methods become more sophisticated for finding certain demographics, we should be able to apply them successfully. Our next set of experiments are meant to demonstrate that the demographics that we found seem reasonable.

Since it would be infeasible to examine the demographics for each Website and determine if they are reasonable, we make the analysis more tractable by first clustering together similar sites. For our clustering, we group together the Websites that share the exact same set of specific demographics. Below are some examples from these clusters that demonstrate reasonable clustering of the sites and their demographics. Each table below defines the demographics at the top of the table, and then lists the clustered Websites where the demographics hold. We note that these demographics represent a super-set in that it is unclear which of them should be disjoined and/or conjoined together.

Table 6 shows a set of recipe Websites. The demographics shared by these Websites are given in the table, and emphasize married bloggers who have full families (e.g., both pets and kids) and these families are likely to be younger (based on the age demographic). Therefore, it seems reasonable that recipe Websites would be interesting to those interested in finding recipes for their families.

Table 7 shows a set of Websites that focus on hard news (as opposed to entertainment news). This group of sites clus-

Table 6: Website Cluster: Recipes

Demographics	
Age 31-40	Is Parent
Is Married	Has Pets
Mentions Parents	Mentions Grandparents
Websites	
thekitchn.com	
bakerella.com	
myrecipes.com	
simplyrecipes.com	
baloon-juice.com	

ters together under the demographics, "Is Grandparent," and "Plays Music." In fact, our "Plays Music," demographic can really be interpreted as "has instrument." For instance, users who have a piano or flute qualify under this demographic. However, it seems reasonable that people old enough to have grandchildren and have an interest in the arts (via musical interest) would be interested in traditional news sources.

Table 7: Website Cluster: News

Demographics	
Is Grandparent	Plays Music
Websites	
wsj.com	
usatoday.com	
huffingtonpost.com	
washingtonpost.com	
latimes.com	
cnn.com	
guardian.com	
foxnews.com	
salon.com	
timesonline.com	
time.com	
dailymail.com	
npr.org	
reuters.com	

Finally, we show that certain demographics appear quite correct. Table 8 shows the next group of Websites, each of which focuses on either video games or comic books. These are traditionally areas of interest to younger men. As expected, these Websites then cluster under the demographics of "In Relationship, Male," and "Mentions Parents." Presumably, these are younger men (as they tend to mention their parents in their blogs), and therefore the assignment of these demographics to this cluster appears correct.

And lastly, the website juicycouture.com is for a clothing brand that caters to teenagers. Indeed, the demographics for this site are teenagers, who tend to talk about their parents in their blog (hopefully to thank them for buying the clothes).

Table 8: Website Cluster: Games & Comics

Demographics
In relationship, Male
Mentions Parents
Websites
supermanhomepage.com
comicscontinuum.com
newyorkcomiccon.com
superherosupplies.com
mangablog.net
fandomania.com
bungie.net
bobshouseofvideogames.com
gamepolitics.com
thedrunkenmoogle.com
ffxiclopedia.org

Table 9: Website Cluster: juicycouture.com

Demographics	
Age 11-20	Mentions Parents
Websites	
juicycouture.com	

4. RELATED WORK

The main goal of our work is to discover demographics of Websites by leveraging outside information, in particular, the blogs that link to them. While there is previous work on trying to discern and use demographic information about users of certain websites (e.g., health websites [4] or e-government sites [8]), these demographics are built from collected surveys. As we state above, this can be problematic due to the low return rate for surveys and their smaller sample size. Therefore, we propose to leverage blogs as a much larger source from which to discover demographics.

The other option is to mine such demographics, leveraging users' past browsing history and possibly some other context such as the site's content, structure, and even a user profile, if it exists. While there is previous research on this task (e.g., [5, 11, 13]), and this work can successfully recommend content for users (e.g., in a personalization setting), it is difficult to turn these mined likes and dislikes of users into demographic groups, because the demographics are latently exploited. That is, there is some latent demographic to the user that the mining discovers (otherwise the personalization would not return items of interest to a user), but it is not explicit enough that the content providers themselves can use it.

Finally, our method leverages "self-referring" facts to define the demographics of our users. There is also previous work on this topic that attempts to discover the demographics of users based on what they write about. In particular, our focus is on blogs, and there is previous work that demonstrates successful text analysis to identify bloggers' native language [7], and their age and gender [12]. As our method is agnostic to types of demographics discovered, we believe these methods are complementary and could yield even more, interesting demographics to include in our approach.

5. CONCLUSIONS

In this paper we presented Blographics, an approach to generating demographics for Websites based upon the blogs that link to them. We demonstrated that we can combine

text analysis methods with network analysis methods to generate reasonable demographics for Websites. These demographics can be used to better align consumers of content with producers of content, used to cluster Websites, and used to gain deeper understanding about bloggers.

Previous work demonstrated that unknown attributes in a social network can be inferred by imputing them through the network [3]. However, it is unclear how we would impute some of our attributes as we only have positive values (e.g., we know someone is a parent, but we might not be able to discover if someone is not a parent). Further, imputing values can introduce noise, and we want our demographics to remain clean. Therefore, in our future work we plan to investigate these issues to allow for imputing attributes in our method.

Further future work involves examining more in-depth network analysis methods, and more sophisticated text analysis methods to develop even deeper demographic profiles. For example, our current approach only considers the one-hop neighborhood around bloggers linking to sites, but ignores the connectedness of this component. That is, most of the bloggers in this neighborhood are likely strongly connected to one another, but there are some outliers that are weakly connected to the set, and likely share less homophily. We plan to investigate how to deal with these issues using more sophisticated network analysis methods.

Lastly, we note that our text analysis focused on frequent patterns in the text and we ignored many of the richer demographic information which is available from bloggers own profile site as well as from their text. For example, a non-trivial number of bloggers referred to their type of abode (apartment, house, condo, etc) as well as region (specific city, town or geographic region). It is clear that there is a ripe research topic on ways to extract deeper demographic information from free text.

6. REFERENCES

[1] M. S. Ackerman and L. F. C. J. Reagle. Privacy in e-commerce: Examining user scenarios and privacy preferences. In *Proceedings of the ACM Conference on Electronic Commerce*. ACM, 1999.

[2] B. Berendt, O. Günther, and S. Spiekermann. Privacy in e-commerce: Stated preferences vs. actual behavior. *Communications of the ACM*, 48(4):101–106, 2005.

[3] S. Bhagat, I. Rozenbaum, and G. Cormode. Applying link-based classification to label blogs. In *Proc. of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 92–101, New York, NY, USA, 2007. ACM.

[4] M. Dutta-Bergman. Trusted online sources of health information: Differences in demographics, health beliefs, and health-information orientation. *J Med Internet Res*, 5, 2003.

[5] M. Eirinaki and M. Vazirgiannis. Web mining for web personalization. *ACM Trans. Internet Technol.*, 3(1):1–27, 2003.

[6] D. L. Hoffman, T. P. Novak, and M. Peralta. Building consumer trust online. *Commun. ACM*, 42(4):80–85, 1999.

[7] M. Koppel, J. Schler, and K. Zigdon. Determining an author's native language by mining a text for errors. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 624–628, New York, NY, USA, 2005. ACM.

[8] H. Li, B. H. Detenber, W. P. Lee, and S. Chia. E-government in singapore: Demographics, usage patterns, and perceptions. *Journal of E-Goverment*, 1:29–54, 2004.

[9] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[10] L. I. Millett, B. Friedman, and E. W. Felten. Cookies and web browser design: toward realizing informed consent online. In *Proc. of CHI*, pages 46–52, 2001.

[11] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Effective personalization based on association rule discovery from web usage data. In *WIDM '01: Proceedings of the 3rd international workshop on Web information and data management*, pages 9–15, New York, NY, USA, 2001. ACM.

[12] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. Effects of age and gender on blogging. In *Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 2006.

[13] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: discovery and applications of usage patterns from web data. *SIGKDD Explor. Newsl.*, 1(2):12–23, 2000.

[14] D. Yates, M. Shute, and D. Rotman. Connecting the dots: When personal information becomes personally identifying on the internet. In *Proceedings of the International Conference Weblogs and Social Media (ICWSM)*, 2010.