A Gender-Centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records

Vanessa Frias-Martinez, Enrique Frias-Martinez and Nuria Oliver
Data Mining and User Modeling Group
Telefonica Research
{vanessa,efm,nuriao}@tid.es

Abstract

The gender divide in the access to technology in developing economies makes gender characterization and automatic gender identification two of the most critical needs for improving cell phone-based services. Gender identification has been typically solved using voice or image processing. However, such techniques cannot be applied to cell phone networks mostly due to privacy concerns. In this paper, we present a study aimed at characterizing and automatically identifying the gender of a cell phone user in a developing economy based on behavioral, social and mobility variables. Our contributions are twofold: (1) understanding the role that gender plays on phone usage, and (2) evaluating common machine learning approaches for gender identification. The analysis was carried out using the encrypted CDRs (Call Detail Records) of approximately $10,000$ users from a developing economy, whose gender was known *a priori*. Our results indicate that behavioral and social variables, including the number of input/output calls and the in degree/out degree of the social network, reveal statistically significant differences between male and female callers. Finally, we propose a new gender identification algorithm that can achieve classification rates of up to 80% when the percentage of predicted instances is reduced.

1. Introduction

The pervasiveness of cell phones in developing economies have made them an ideal platform for providing many services centered on improving local living conditions (Inter American Development Bank 2009). For example, ZMQ and Grameen Foundation use SMSs to inform women about prenatal care in India and Ghana respectively (ZMQ 2008; Grameen 2009). Similarly, Project Masiluleke offers HIV/AIDS and TB education, as well as awareness programs for men and women in South Africa via cell phones. This initiative resulted in an increase of 350% in the volume of calls to their hotline (Masiluleke 2008). While some of these programs offer gender-neutral solutions, it is clear that many applications would be implemented most effectively with prior knowledge of the gender of the person at the receiving end of the service. Additionally, the gender divide in the access to technology – that makes women more susceptible to technological illiteracy (Diga 2008)– together with country-dependent social and cultural factors drive the need to develop personalized gender-based services. Hence, it is both through a better understanding of gender-related differences in the use of technology (gender characterization) and the correct identification of the gender of specific cell phone users (gender identification) that cell phone-based services can be improved in developing economies.

Gender characterization has been investigated by the human-computer interaction (HCI) and psychological communities in both developed and developing regions. Female cell phone users in the UK were found to be more comfortable than males making or receiving personal calls in different social contexts (Turner, Love, and Howell 2008). Similarly, research has demonstrated that males in West Africa tend to use cell phones for job-related tasks as opposed to females who tend to use them for personal calls (Huyer et al. 2006). Although these studies offer important insights that can be helpful towards gender characterization, such results are typically based on questionnaires applied to a limited amount of individuals.

The topic of gender identification (or gender classification) has been extensively studied by the computer vision and speech processing communities (Shue and Iseli 2008), (Yang, Li, and Ai 2006). However, these approaches for gender identification algorithms require access to the content of private conversations or private images, which in the context of cell phone networks is not feasible due to privacy concerns.

The goal of this paper is to characterize and automatically identify gender using only behavioral, social and mobility information obtained from Call Detail Records (CDRs) of a developing economy. In the context of developing economies, this study is particularly relevant since pre-paid clients for whom there is no specific gender information account for the large majority of cell-phone users. Our approach reduces the privacy concerns, by avoiding the use of actual conversations, and allows us to model large populations without the need to deploy questionnaires, since millions of calls with behavioral information are available in the CDRs. CDRs have been already used as a source of information for understanding and modeling user behavior (Dasgupta, Singh, and Viswanathan 2008; Nanavati and Gurumurthy 2006). However, and to the best of our knowledge, no specific studies on gender characterization and identification from CDR data have been carried out to date. Hence, the novel contributions of this paper are:

- A gender-centric analysis of the calling behavior of cell phone users in a developing economy. Female and male cell phone usage has been characterized as a function of three types of variables: (1) behavioral, *e.g.,* the number and duration of the calls; (2) social, *e.g.,* in and out degrees of a user's social network; and (3) mobility, *e.g.,* the total distance traveled by the user.

- A novel semi-supervised clustering algorithm that captures typical female and male calling behaviors. Our approach yields classification accuracies between $70\% - 80\%$ with coverages (percentage of instances that can be classified) of up to 10%.

2. Related Work

A study of mobile phone usage in Uganda has revealed a clear gender imbalance (Diga 2008). In particular, Diga has shown that there exists an unequal partner control and usage of the cell phone, specially inclined towards male ownership. Comparable results have been also obtained by Huyer *et al.* (2006), whose analysis examined the use of cell phones and internet in West Africa. These authors also found that men tend to use cell phones for professional or work-related tasks, while females favor social and personal calls. A recent study in India, Mozambique and Tanzania concluded that males use cell phones with a higher frequency than females, probably because of social norms and financial considerations (Souter et al. 2005). In addition, the authors observed that men appear to regard cell phones more highly than women, particularly for business activities.

Intriguingly, other studies have shown that the gender gap in cell phone usage is narrowing, with men and women reporting nearly identical calling behaviors (DeBaillon and Rockwell 2005). In a gender-based study of cell phone usage in Pakistan, India, Sri Lanka, Philipinnes and Thailand, Zainudeen et al. (2008) showed that for all countries, except for Pakistan, women have similar call frequencies, call destinations and call durations as men. Similarly, Wilska (2003) reported that in developed (advanced) economies, the traditional gender division regarding cell phone usage has disappeared among young people, where both females and males show typical *male attitudes* such as technology enthusiasm or even addiction.

Taken together, previous research works highlight the existence of gender-based differences as well as similarities in calling behaviors across developing economies. Nevertheless, such studies typically come from the field of psychology based on results that are usually derived from a limited number of personal interviews and/or questionnaires. In this paper, we aim at overcoming these limitations by using CDR data collected from a large and diverse population.

3. Dataset Description

Cell phone Call Detail Records (CDRs) from a developing economy [1] were collected from a major carrier for a period of three months. Each CDR contains the encrypted cell phone numbers of caller and callee, the date and time of the call, the duration of the call and the initial and final location of the caller while making the call. The caller location is approximated by the geographical position of the cell tower that handled the call. The dataset contains approximately 2 million calls that correspond to around $10,000$ unique cell phone numbers: approximately $5,000$ females and $5,000$ male clients. Only users that maintained a phone number with the carrier throughout the three months of study are considered. The cell phone usage expenses incurred over the period of study and the gender associated to each encrypted number (user) were also available.

4. Description of Characterization Variables

Behavioral Variables

Behavioral variables characterize the behavior of a user in terms of calling consumption: number of calls, duration and expenses generated by those calls.

- *Number of Calls*: We consider both incoming calls *i.e.,* the total number of calls received by user j during a period of D days, as well as outgoing calls *i.e.,* total number of calls made by user j during a period of D days.
- *Average Duration of Calls*: The incoming average duration for user j is calculated as the duration of all the incoming calls divided by the total number of incoming calls received by the user. Similarly, we define an outgoing average duration taking into account all the outgoing calls placed by the user.
- *Expenses*: The expenses of user j correspond to the total amount of money spent by the user in phone calls over a period of D days.

Social Variables

Social variables characterize the social network of a user based on her/his use of the cell phone.

- *In/Out Degree of the Social Network*: The *in degree* for user j is given by the number of different phone numbers that called that user over a period of D days. This variable is calculated as the cardinality of the union of all incoming unique phone numbers for user j. Analogously, the *out degree* is defined by amount of distinct phone numbers contacted by user j.
- *Degree of the Social Network*: The degree of the social network corresponds to the number of *unique* phone numbers that have either contacted or been contacted by user j.

[1] Based on the 2009 International Monetary Fund (IMF) World Economic Outlook Country Classification www.imf.org/external/pubs/ft/weo/2009/02/weodata/groups.htm

Distribution	PowerLaw Fit (a,xmin)	LogNormal Fit (μ,ω)	Exponential Fit (λ)
(Number of Output Calls, Female)	(3.33,338)*	(4.1395,1.2695)	124.24
(Number of Output Calls, Male)	(3.01,226)*	(3.922,1.27)	101.14
(Average Duration of Output Calls, Female)	(3.65,77.8)	(4.56,0.94)*	160.33
(Average Duration of Output Calls, Male)	(2.1,49)	(4.42,0.91)*	138.49
(Expenses, Female)	(4.74,290.0)	(3.81,1.15)	78.0*
(Expenses, Male)	(3.94,218)	(3.65,1.18)	69.53*
(Degree, Female)	(3.5,38)*	(2.65,0.89	20.38
(Degree, Male)	(3.5,36)*	(2.559,0.907)	18.77
(Route Distance, Female)	(1.79,1.23)	(0.1695,1.95)*	5.087
(Route Distance, Male)	(1.75,1.27)	(0.1205,2.16)*	5.68

Table 1: Fitting parameters for each $(variable, gender)$ distribution using Clauset's fitting algorithm and Matlab statistical toolbox. The fitting parameters are: exponent a and value of x where fitting starts $(xmin)$ for the power-law; mu and $omega$ for the log-normal fit; and exponent λ for the exponential. The * represents the best fit which was obtained using the Kolmogorov-Smirnov goodness-of-fit statistic.

Mobility Variables

Mobility variables are used to characterize the movement of cell phone subscribers from the information available in the CDRs, *i.e.,* the location of the cell towers that the mobile phone connected to when a call started and ended.

- *Route Distance*: The route distance is the distance traveled by user j *between* consecutive calls. For a pair of calls, it is computed as the distance between the coordinates (latitude,longitude) of the tower where the first call ended and the coordinates (latitude,longitude) of the tower where the second call started. This distance approximates the route that the user has traveled between each pair of calls. As such, it gives a coarse approximation of the geographical mobility of the user. The final route distance for user j is computed as the sum of the route distance of all calls for that user, divided by her/his total number of calls.

5. Gender Characterization

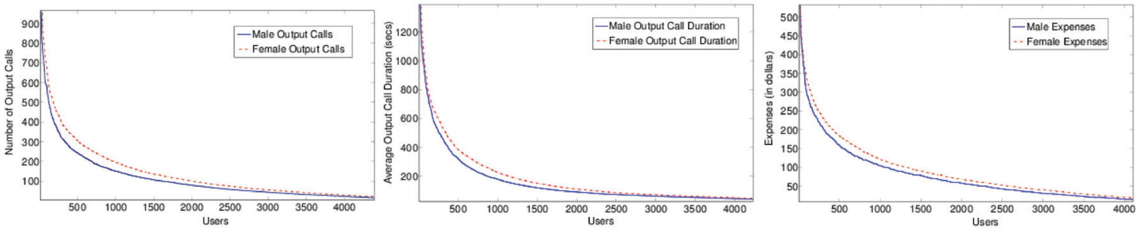
In order to understand the calling behavior of females and males in the CDR dataset, we computed the distributions for each pair $(variable, gender)$ where $variable$ represents one of the behavioral, social or mobility variables described in the previous section, and $gender$ is either female or male. Each distribution contains the set of values that the variable acquires for all female or male individuals within the dataset. For instance, the distribution $(Number\ of\ Output\ Calls, Female)$ contains approximately $5,000$ entries that correspond to the total number of outgoing calls made by each female in the dataset throughout the three months of CDRs.

Calling behaviors have been typically associated with power-law, log-normal or exponential distributions (Clauset, Shalizi, and Newman 2009). In an attempt to understand the nature of our gender-based distributions, we applied fitting algorithms with the three most common distributions in the literature (power-law, log-normal and exponential) to each of the $(variable, gender)$ distributions.

Table 1 shows the fitting parameters of the power-law, log-normal and exponential functions for each $(variable, gender)$ distribution. In the case of bidirectional variables such as incoming or outgoing number of calls, incoming or outgoing duration and in or out degree, we only report the results in the outgoing direction. Comparable results were found in the incoming case. The marked entries in the Table denote the fittings with the best fit, based on the Kolmogorov-Smirnov goodness-of-fit statistic. Our findings include that the $(number\ of\ calls, gender)$ and $(degree, gender)$ distributions for both female and male are best fitted by a power-law distribution. Conversely, the $(average\ duration\ of\ the\ calls, gender)$ and $(route\ distance, gender)$ distributions follow more closely log-normal distributions. Finally, a exponential fit works best for the $(expenses, gender)$ distribution. Thus, these results confirm that the gender-based distributions in our CDR dataset follow similar trends to the ones already reported in the literature (Clauset, Shalizi, and Newman 2009).

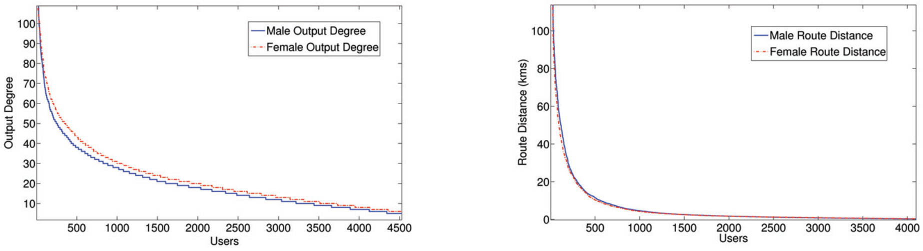
Next, we analyze in detail each of the computed $(variable, gender)$ distributions. Figures 1(a), 1(b) and 1(c) depict the ranked distributions for each pair $(variable, gender)$ for the behavioral variables. Only the outgoing direction is plotted for bidirectional variables. Figure 1(a) shows the number of calls made by each user within the female and male distributions. We find that nearly 10% of females made more than 300 calls over the three months. In contrast, the male population shows less activity across all users. On the tail of the distribution, approximately 60% of the female population and 70% of the male population made fewer than 100 phone calls. We observe a similar pattern for the average duration of the calls (see Figure 1(b)), where around 10% of the females had an average call duration greater than 400s as opposed to their male counterparts with a duration of 365s. In terms of expenses (Figure 1(c)), 7% of the females spent more than $\$200$ in three months, whereas only 4% of the males had similar expenses. The tail of Figure 1(c) reflects that on average males spend less than females. In general, we observe that *females have higher usage levels than males for the behavioral variables.*

Similarly, *the female distributions have higher values than the male distributions in the social variables.* In Figure 2(a) we observe that approximately 7% of females have an output social degree of at least 50, whereas only 5% of



(a) (b) (c)

Figure 1: Female and Male ranked distributions for the behavioral variables: (a) Number of Output Calls, (b) Average Output Duration, and (c) Expenses. The number of output calls, the average output durations and the expenses are computed for each individual of the CDR dataset.



(a) (b)

Figure 2: Female and Male ranked distributions for the social and mobility variables: (a) Output Degree and (b) Route Distance. The output degree and route distances are computed for each individual of the CDR dataset.

the males achieve similar values. Again, the tail shows that around 20% of the females and 35% of the males have an output degree smaller than 10. Similar trends were observed for the variables input degree and degree.

Finally, *in terms of route-distance, females and males have quite similar behaviors*. Figure 2(b) shows that approximately 9% of either population travels on average more than 10kms in between consecutive calls. We also observe a heavy tailed behavior whereby around 68% of the population travels less than 3kms across calls.

Although this study is limited to one developing economy, similar analyses should assist developers and international organizations in the design of future gender-based cell phone services. For instance, our results show that in the country under analysis women tend to have a large number of social contacts. Therefore, new cell phone services for these women should favor social networking, and take advantage of it when implementing new solutions.

Statistical Differences in the Calling Behavior

The previous analysis showed preliminary evidence suggesting gender-based differences in the way individuals use their mobile phones to communicate with others. In order to test the statistical significance of our results, we conducted a series of statistical tests for all behavioral, social and mobility variables. In particular, we computed a t-test (the Central Limit Theorem holds under the assumption that our distribution is sufficiently large) and a two-sample Kolmogorov-Smirnov (K-S) test for each pair of $\{(variable, female), (variable, male)\}$ distributions.

We found that the differences between the female and male distributions are statistically significant (with $p < 0.01$) for the *behavioral* and *social* variables. However, the null hypothesis can not be rejected for the female and male distributions representing the mobility variable of route distance which indicates that *females and males exhibit similar mobility behavior*. Finally, in order to complement our analysis, whenever the *null hypothesis* was rejected we also ran the one-sided (right-sided) version of the Student's t-test. In such tests, we found that the *mean of the female distributions is always larger than the mean of the male distributions*, thus corroborating the trend observed in our analysis.

6. Gender Identification from Call Data

In order to characterize each individual, we modeled the calling behavior b_j for an individual j as a function of the following variables: number of input calls and output calls, average duration of input and output calls, expenses, and degree of the social network. The values for each variable were calculated over the total period of three months. The remaining variables were not considered either because they did not reveal statistically significant differences across gender, or because they had a high degree of cross-correlation.

In order to train the models, we divided the calling behavior data into a training set (70% of data) and a testing set (30%). Hence, the training set consisted of approximately $7,000$ calling behaviors ($3,500$ female-labeled and $3,500$ male-labeled –randomly selected– calling behaviors).

In turn, the testing set contained the remaining $3,000$ calling behaviors ($1,500$ female-labeled and $1,500$ male-labeled). Random selections of different training and testing sets were repeated 60 times for each experiment. The results reported correspond to the average values, however little variation was detected across runs.

Gender Classification Algorithm

We first evaluated the accuracy of two well known supervised classification techniques: Support Vector Machines (SVMs) and Random Forests. However, the classification rates (54.2% for SVMs and 56% for Random Forests with boosting) were very low when compared to the 50% random probability of identifying the gender correctly. Thus, in an attempt to improve the results from the supervised classification, we present a semi-supervised clustering algorithm that improves the accuracy of the classification by reducing the percentage of predicted instances.

The proposed algorithm consists of a training and a testing phase. Initially, the training phase of the algorithm starts by applying *k-means* clustering to the training data for a specific value of *k* (Hartigan and Wong 1979). In our case, *k-means* will distribute the calling behaviors b_j in the training set into k clusters. The algorithm then labels each resulting cluster c with a female or male tag, based on a minimum percentage requirement p of female or male-labeled instances within the cluster. Higher values of p guarantee more precise definitions of female or male behavior since each cluster would consist of a large number of instances of a particular gender. In contrast, lower values of p would produce fuzzier behavioral definitions since the clusters contain more even mixtures of female and male instances. It is important to note that the algorithm does not label clusters that fail to reach the minimum percentage requirement p. The final gender classifier is built using only clusters that have been labeled female or male by the algorithm. Subsequently, these clusters are the model of what constitutes female and male behavior for the gender classification algorithm.

For each labeled cluster c, a *radius* ($radius(c)$) is computed. This radius is given by the maximum distance $maxDistance$ between a calling behavior b_j in cluster c and the centroid of the cluster ($centroid(c)$). These radii are used in the testing phase as a measure of similarity of the testing instances to each of the clusters. The pseudocode for the training phase is shown in *Pseudocode 1*.

During the testing phase, the algorithm finds the closest labeled cluster c to each calling behavior b_j in the test set. If the distance between b_j and the centroid of cluster c ($centroid(c)$) is less or equal than the cluster's radius ($radius(c)$), the calling behavior b_j is assigned the gender label $gender(c)$ of that cluster. Otherwise, no gender prediction is provided due to the lack of sufficient gender certainty in the classification algorithm (do-not-know label). Therefore, the classification algorithm sacrifices the coverage of the classification (percentage of classified instances) in order to improve its accuracy (percentage of correctly classified instances).

Pseudocode 1 Training Phase.

```
clusterSet = kmeans(k,$\{b_j\}$)
for each cluster $c$ in $clusterSet$ do
  if ((males($c$)/ (females($c$)+males($c$)) $> p$) then
    gender($c$) = male
  else if ((females($c$) / (females($c$)+males($c$)) $> p$) then
    gender($C$)=female
  else
    gender($c$) = ””
  end if
end for
for each labelled cluster $c$ in $clusterSet$ do
  maxDistance = 0
  for each instance $s$ in $c$ do
    if (distance($s$,centroid($c$)) $>$ maxDistance) then
      maxDistance = distance($s$,centroid($c$))
    end if
  end for
  radius($c$) = maxDistance
end for
```

Experimental Results

Figure 3 shows the accuracy and coverage of the proposed gender classification algorithm for various values of k: 10, 35, 50 and 100 clusters. Additional k values were considered, but we only discuss a few that show the general trend. For each value of k, the $(accuracy, coverage)$ pairs are obtained by progressively decreasing the minimum percentage requirement p. As expected, for all values of k, the lower the coverage of the classifier, the higher the accuracy rates and *vice versa*. This trend derives from the fact that as the value of p decreases, more clusters are assigned a gender label and the coverage increases. In addition, higher values of p reduce the coverage but greatly increase the accuracy of the classification since the requirements for female *vs* male behaviors are more strict.

The proposed algorithm allows us to determine the accuracy or the coverage needed for a particular application. For example, accuracies of around 80% are obtained with coverage rates of 3% ($k = 50$). In contrast, lowering the accuracy knob to 70% yields a coverage rate increase to approximately 12%. Although these coverage rates may seem low, being able to automatically classify the gender of 12% of a population of one million with 70% accuracy translates into accurate gender classification for as many as $120,000$ individuals. This is specially relevant in the context of developing economies where a vast majority has pre-paid cell phones for whom the gender is not known.

In order to gain further insights into the gender classification algorithm, we selected the set of labeled clusters that provided the highest accuracy (acc) for each value of k ({k=10,acc=63%}, {k=35,acc=71%}, {k=50,acc=80%} and {k=100,acc=77%}). Interestingly, for all values of k the labeled clusters that produced the highest accuracy were all tagged as female. This indicates that the gender classification algorithm produces its highest accuracies when predicting female-only calling behavior. An analysis of the instances that were correctly predicted as females with the highest accuracy, showed that these particular females had

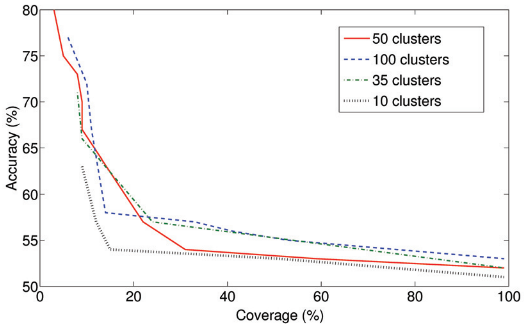


Figure 3: Accuracy versus Coverage for sets of clusters with size: 10, 35, 50 and 100 behavioral clusters.

calling behaviors well above the average in our testing set, specifically for the social variables. This finding may be interpreted as an above average female group, with high social degrees, that peaks prominently over the bulk of the population of regular callers.

7. Conclusions and Future Work

To the best of our knowledge, we have presented the first study aimed at characterizing gender based on calling behavior obtained using CDR data from a developing economy. We have shown that female and male behavioral and social variables follow distributions which are statistically different ($p < 0.01$). Conversely, we have not found statistically significant differences between female and male mobility. In addition, we have presented a semi-supervised gender classification algorithm that can predict gender with 80% accuracy when the coverage is around 3% while 70% accuracy is reached for a coverage of around 12%. In particular, we have detected the existence of a small but well defined set of female users with a social degree well above average.

In closing, we note that it is crucial to take into account these differences in calling behavior across gender when designing gender-specific cell phone services. A key open research question, left for future work, are the implications that the findings of this paper have for the design of mobile services.

Future work will study whether individuals of the same gender display similar behavioral patterns across different countries. In addition, we plan to study and characterize additional factors that play a role in the use of cell phones including income, age, and level of education.

References

Clauset, A.; Shalizi, C.; and Newman, M. 2009. Powerlaw distributions in empirical data. *SIAM Review*.

Dasgupta, K.; Singh, R.; and Viswanathan, B. 2008. Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of EDBT*, 668–677.

DeBaillon, L., and Rockwell, P. 2005. Gender and student-status differences in cellular phone use. *International Journal of Mobile Communications* 3(1):82–98.

Diga, K. 2008. Technology spending patterns and poverty level change among households in Uganda. In *Workshop on the Role of Mobile Technologies in Fostering Social Development*.

Grameen. 2009. Mobile technology for community health (MoTeCH). www.grameenfoundation.applab.org/section/ghana-health-worker-project.

Hartigan, J., and Wong, M. 1979. A k-means clustering algorithm. *Applied Statistics* 28(1):100–108.

Huyer, S.; Hafkin, N.; Ertl, H.; and Dryburg, H. 2006. Women in the information society. In *From the Digital Divide to Digital Opportunities*.

Inter. American Development Bank. 2009. Mobile citizen, solutions at hand. www.mobilecitizen.bidinnovation.org/en.

Masiluleke. 2008. HIV and TB in South Africa. www.poptech.org/project_m_the_solution/.

Nanavati, A., and Gurumurthy, S. 2006. On the structural properties of massive telecom call graphs: Findings and implications. In *Proceedings of the 15th ACM CIKM*, 435–444.

Shue, Y., and Iseli, M. 2008. The role of voice source measures on automatic gender classification. In *IEEE ICASSP*.

Souter, D.; Scott, N.; Garforth, C.; Jain, R.; Mascararenhas, O.; and McKemey, K. 2005. The economic impact of telecommunications on rural livelihoods and poverty reduction. In *Commonwealth Telecommunications Organization for UK Department for International Development*.

Turner, M.; Love, S.; and Howell, M. 2008. Understanding emotions experienced when using mobile phone in public: The social usability of mobile (cellular) phones. *Telematics and Informatics* 25(3):201–215.

Wilska, T. 2003. Mobile phone use as part of young people's consumption styles. *Journal of Consumer Policy* 26(4):441–463.

Yang, Z.; Li, M.; and Ai, H. 2006. An experimental study on automatic face gender classification. In *ICPR*, 1099–1102.

Zainudeen, A.; Iqbal, T.; Samarajiva, R.; and Ratnadiwakara, D. 2008. Who's got the phone? the gendered use of telephones at the BOP. In *Annual meeting of the International Communication Association*.

ZMQ. 2008. mHealth for development: A UN and Vodafone Foundation report. www.unfoundation.org/global-issues/technology/mhealth-report.html.