Challenges and Opportunities in Data Mining Contact Lists for Inferring Relationships

Jason Wiese, Jason I. Hong, John Zimmerman
Carnegie Mellon University
{jwwiese, jasonh, johnz}@cs.cmu.edu

ABSTRACT
The smartphone contact list has the potential to be a valuable source of data about personal relationships. To understand how we might data mine the information that people store in their contact lists, we collected the contact lists of 54 participants. Initially we found that the majority of contact list features were unused. However, a further examination of the "name" field revealed a broad variety of contact-naming behaviors. We observed contact "name" fields that included affiliations, relationship role labels, multiple names, phone types, and references to companies / services / places. People's appropriation and usage of contact lists have implications for automated attempts to merge or mine contact lists that assume people use the features and structure of the contact list tool as intended. They also offer new opportunities for data mining to better describe relationships between users and their contacts.

Categories and Subject Descriptors
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Keywords
Mobile Contact List, Inferring Social Context

INTRODUCTION
For many, smartphones have become the nexus for communication and coordination, relying on the contact list as a critical tool. These lists store not only names and phone numbers, but also home and work address, website, display name, organizational affiliation, a photo, and other means of contact (e.g., email, chat, Skype). Individual contacts can be selected as "favorites," grouped into categories, and linked with social networking profiles.

Smartphones in general and the contact list in particular have the potential to be valuable sources of data about personal relationships because of the information they store. For example, both the content and the metadata from phone calls, emails, and SMS messages can be used to model social relationships, facilitate online sharing [8], personalize interfaces, or contextualize communication.

UbiComp '14, September 13 – 17, 2014, Seattle, WA, USA

ACM 978-1-4503-2968-2/14/09...$15.00.
http://dx.doi.org/10.1145/2632048.2632108

To understand how we might data mine the information that people store in their contact lists, we collected the contact lists of 54 participants, containing 35,599 contacts. 67% of the contact entries that we collected contained either no contact information, or only an email address. Most of the remaining 33% of contacts only contained one piece of information, usually a phone number. The majority of contact list features were unused.

Despite the apparent lack of information contained in these lists, a deeper exploration of the content uncovered more subtle structures within the data. Analysis of the contact name field yielded twelve distinct and unexpected naming strategies, including affiliation (*Pat (Neighbor)*), familial roles (*Aunt Joan*), phone type (*Mom at Home*), and calls to avoid (*Do Not Answer*).

These observations of real-life contact list usage point to implications for mining the data held within them. Automated attempts to data mine users' contact lists or automatically merge contacts from multiple lists will likely produce poor results if they assume people use the features and structure of the contact list tool as intended. On the other hand, the structure of a user's contact naming behaviors offers new opportunities for data mining to better describe relationships between the user and her contacts.

RELATED WORK
The literature on personal information management (PIM) focuses on the broad set of data that relates directly to the user (such as email, documents, and personal notes). This can include contact information; yet, smartphone contact lists have not been a focus of this work. One particularly relevant PIM topic is file naming. Jones [4] offers several important insights for naming files:

1. It is easier to include all of the metadata in the file title than to enter metadata in structured fields
2. Artifacts often have one key property: papers might be named by the author's last name, while photographs might be named by date and time taken.

This PIM work offers insights for our observation of contact naming conventions. Our work also exposes the complex nature of the "key properties" identified by Jones. Unlike the structured examples above, key properties for contact lists seem to depend on factors specific to the user, the contact, and the context of their relationship.

Feature	*#Entries (out of 35k)*	*# Particp (out of 54)*	*Median per particp*	*Notes*
Events	669	23	4	Almost entirely birthdays, includes some anniversaries
IM	286	13	2	Instant messenger screen names
Address	1091	26	7.5	Street address information
Notes (all)	1906	31	7	Mostly syncing/auto-populated data (signified by a consistent XML structure), including Facebook profile IDs, like addresses, names
Notes (manual)	188	24	3	Various manually-entered text
Organizations	612	23	5	Company and/or job title
Relations	191	4	12.5	Mostly child/spouse data
Websites	1951	36	31	Mostly Facebook and Google profiles. Excluding those, 44 remain.

Table 1. A summary of the usage of contact list fields for our dataset. Less than half of participants used most of these. Furthermore, many of the fields are auto-populated and used inconsistently across different participants.

Whittaker *et al.* [7] investigated the many challenges in managing contact information. This work, done well before the advent of smartphones, noted people's need to maintain a *manageable* list. Other work has examined contact list entries from different users and used a similarity algorithm to identify contacts that are likely to be the same person [2]. This could automate the updating of contact lists, and could resolve duplicates by leveraging the information contained in other users' contact lists.

Work by Min *et al.* leveraged several features from the smartphone to classify contacts as *family*, *work*, or *social* [6]. These features included contact entry similarity to the user, which parts of the entry were filled in, and if a contact was starred. The value of the contact list features was limited by its many empty fields. Our work provides insight into some of these issues, and uncovers patterns in contact naming that could improve classification efforts.

Bentley *et al.* discuss the process of creating an app that combines the contact list with social media [1]. They found that people tend to think about their contacts in clusters, similar to other recent work [3]. Their studies also revealed that users want control over how contacts are combined when syncing across multiple sources. Our work provides support for these findings through insight gleaned from the data within the contact list. We also provide a detailed description of many different ways contact lists are used.

RESULTS

We collected contact lists from 51 Android users (35 female), recruited using Craigslist across the U.S. and from two online bulletin boards. All participants had been using their Android phone for at least six months prior to the study and had a variety of Android phone models. Participants varied in age (range: 19-51, mean: 28). The number of contact records per participant varied dramatically (range: 10-3,237, mean: 659, median: 514). Each contact entry has a name, and some have additional information. 11% of entries contained a name and no other information. 56% had only a name and an email address.

On average, 42% of a participant's contacts had at least one phone number (range: 0-750, mean: 182, median: 132); 67% had at least one email address (range: 2-1748, mean: 489, median: 387); and 13% had both a phone number and an email address (range: 0-573, mean: 78, median: 19). Most phone numbers were labeled as mobile (72.6%), followed by home (16.0%), work (6.0%), and other (5.4%). The shift away from landlines to mobile and the default settings of most Android devices to label new numbers as mobile may explain the high number of mobiles listed.

Across the dataset, the average number of pieces of information associated with a contact was 1.1. Even excluding the 60% of contacts who had no information, or just an email address, we still found that contacts had very little information associated with them (range: 1-6, mean: 1.7, median: 1). Table 1 contains a summary of the usage of other contact list fields. The overwhelming majority of entries contain none of this additional information. Furthermore, much of the information that was included, such as entries for *websites* and *notes,* did not appear to belong to those fields: they seem to be profile keys used by contact-syncing with social network sites. In all, 2,253 contacts had evidence in the website or note field that they were automatically synced from a social networking site, however excluding these contacts did not affect the general magnitude of the results (e.g. contacts per participant: range: 10-2,832, mean:618, median: 501).

At a high level, these data indicate that mining contact lists for context about relationships cannot simply rely on the structured data fields of the contact list.

Contact Naming Conventions

Many contact name entries deviated from the *Firstname Lastname* convention. To investigate further, we sampled 15% of the entries from across participants to examine manually, looking for unexpected and unusual uses of the name fields. Using an open-coding process, we found twelve categories of naming strategies that deviate from the *Firstname Lastname* convention. Next, we followed a closed-coding process with an additional 10% of entries, which allowed us to validate the categories and their occurrence over more than one participant. Our goal was to understand how common these naming strategies are, and to see if we could develop insights into what we might be able to mine from the name field of contact entries.

We grouped these twelve strategies into four categories: socially defined; companies, services, and places; tasks; and

other. Then we generated a set of simple rules and used pattern matching against the entire dataset to roughly count the number of times participants employed each strategy. We excluded contacts with an email address in the name field, likely an error from merging two lists where an individual record only had an email address. All identifiable information is anonymized, but the structure is maintained.

Socially Defined Contacts

These naming conventions include people mentioned by name, with a fair amount of variation within this category:

First name without a last name (e.g., *Tommy, Sandy, Pat*): 9.4% of all contacts were listed with a completed first name field that did not contain any spaces, and an empty last name field. Having just a first name may indicate familiarity (i.e. there is only one "Tommy" that this could possibly be, a last name in unnecessary).

Honorific included (e.g., *Mrs. Greenman, Mr. Joseph, Officer Gene*): Perhaps indicating formality, distance, or status, this occurred in 0.1% of all contacts.

Group as a single social unit (e.g., *Mel & Cindy Tanner, Mom and Dad*): Contact lists are designed for one-person per record. However, we found *&* and *and* occurred in 5% of all contact entries, a workaround that suggests the participant views a set of people as a single social entity.

People with more than a first and last name (e.g., *Michael Brien Daniels, Sidney G Major, Jr.*): These names may increase confusion in automated systems, particularly for merging contacts across multiple systems.

Family Role (e.g., *Aunt Joan, Grandpa Jim, Mom*): 0.6% of all contacts contained one of the following role labels: *mom, dad, aunt, uncle, grandma, grandpa,* or *cousin*. These roles labels can improve data mining, but can also cause problems for automated systems since one person may be listed with different names across contact lists.

Relationship Context (e.g., *Kaitlyn (Peg's Friend), Shelly's Dad, steve from work, person From Oakland, dane from gym, Chris Group Ga Dawgs Fan, Julia Janson Sells Mary Kay, Jenn From Floor 8, Sandy (next door), Zack New York, Brynn (Meetup)*): This strategy uses affiliation or relationship provenance as a critical component. It suggests that this additional context is needed to distinguish this *Steve* from the other *Steves* in the list. Further, it implies that people search their lists by first name and then use the context to make their selection. At least 17 contacts were associated with a friend of a friend. Additionally, 16 contacts contained *from,* likely indicating social context, and at least 4 contacts were indicated as being neighbors.

Phone Type (e.g.: *Mom at Home, Ranjeet Cell Phone, JBs New Phone*): 0.9% of all contacts contained the word *phone, home, cell* or *mobile*, suggesting multiple contact records that could be merged. Though there is a specific field for *phone type*, users were unaware of it or ignored it, disambiguating phone type in the name field instead.

Companies, Services, and Places

These contact entries represent companies, physical locations, or other phone-based items.

Person with Place or Company Affiliation (e.g., *Ariel Credit Restoration, Mario Meyer -Senior Cab*): This strategy favors affiliation as a critical component in defining a contact. This functions as a reminder of a service agent's name. Though Android contact lists have explicit fields for company name and job title, these fields are used infrequently. 23 of our 54 participants had at least one entry that used the *company name* field, but only 15 participants had more than two entries with company name filled in, and only 2 participants had 15 or more companies listed.

Place/Company names (e.g., *Jefferson Middle School, Klein's Pharmacy, Rizzos Pizza*): This strategy indicates a place instead of an individual. It captures places that a participant might frequently contact, such as her child's school, or services she repeatedly uses, like car repair or a takeout restaurant. Unlike the *friend of friend*, these entries do not show linked affiliations, (e.g., my daughter's school) but instead use the entity's name. *Home* was common (30% of participants). It may indicate the participant's landline or a landline for their childhood home.

Callers to be Avoided (e.g., *Do Not Answer, Telemarketer, Law Office – Do Not Answer*): This strategy highlights a set of callers that the user wants to avoid. Interestingly, the user has taken the step to save the number, indicating that the user expects these people to call again. Also, many of these contacts do not contain any additional context: 7 out of 9 simply say *Don't/Do Not Answer*.

Task Names

(e.g., *Check balance, Check Minutes, Check TextUsage, Paypal Bal*): Several participants had entries indicating an information service accessible from their phone. The contact list is not designed to support this type of entity, so participants used the first name and last name fields to hold this information. Also, many of these tasks are now supported directly within the Android operating system, as well as by apps on the phone, obviating the need for these numbers. It is possible that these contact entries were automatically imported from a previous (non-smart) phone.

Other Contacts

Additional contact strategies did not support strong interpretations. For example, 20% of contact entries contained only an email address, with no first or last name. Other entries included ambiguous name labels, which may provide meaning to the participant but would not provide meaning for automated systems (e.g., *Who are you??, 4, Q, Corn, Eclipse, 555-867-5309, souvenir, Unknown*).

DISCUSSION

There is a discrepancy between the way that the contact list tool was designed to be used and its real-life usage. While the contact list can store many different kinds of structured information, the vast majority of that capability remains

unused, instead storing this information in the name field of the contact entry. To understand this, we need to consider real-life use cases for contact list data. For example:

1. The user needs to identify an incoming caller
2. The user wants to find a specific person in their contact list (e.g. to contact that person)
3. The user wants to find a particular piece of information about a contact entry (e.g., birthday)

The current design of contact lists fits the third use case well. However, the first two use cases (recalling who a contact is or retrieving a particular contact) expose the problem. Despite the broad variety of contact and dialer applications that are bundled with phones or available in app stores, these applications consistently assume that there is one key property to a contact entry: the contact's name. Evidence of this assumption pervades these interfaces: Contacts are listed alphabetically by name, searching for a contact will only query the name fields, SMS messaging apps list contacts by their name, and the caller ID screen that pops up for an incoming call displays the caller's name (or their phone number if no name is in the contact entry).

By contrast, this work provides evidence that the key property for identifying a contact is not necessarily the contact's name. In cases where the user needs information other than the contact's name in order to identify an incoming caller or search for that entry, the information must be stored in the name field. In a way the current approach works: people are able to manage and retrieve their contacts. On the other hand, the contact list has the potential to provide underlying support for applications to understand and design for the social complexity that is inherent in our everyday lives. The potential to support social complexity in user interfaces hinges on our ability to capture that complexity. In this regard, the current approach fails: social complexity is not captured at all, or it is only captured in the name field and is unused by applications.

Beyond a more comprehensive contact list redesign, our work also suggests some new opportunities. It seems feasible to mine contact name fields for some inferences about the user, the contact, or their relationship. This approach has applications for social science research, for personalizing user interfaces, and as a new kind of context that can be used for communication tools.

Some contact naming strategies are likely to be common patterns. For example, *Mom*, *Grandpa*, or *Uncle David* indicate family. Even if this only applies to a small number of contacts per user, it is likely to be robust across many users. Furthermore, mining these can help reveal many of the social roles the owner enacts. Affiliations (e.g., *Kathlyn, Peg's Friend*) show provenance and indicate a stronger tie to the affiliated person than to the specific contact.

In some ways, the automatic merging and adding of contact information (when it works) from services like Facebook and LinkedIn actually results in a loss of data mining potential for contact lists. The user's intention in manually adding information to a contact entry is a valuable signal that indicates user effort. If large amounts of data are automatically added and automated merging actually succeeds, the signal of user intention is very easily lost. The design of contact list data structures can very easily make these differences explicit. Other simple timestamp fields such as *created at* and *updated at* would also be easy to add and could be valuable sources of context for some data mining tasks (e.g., how long you have known this person).

LIMITATIONS AND FUTURE WORK

This work identifies 12 categories of naming conventions that break the assumed contact list convention of *Firstname Lastname*. While these represent a broad set of naming conventions, more might exist. This list is a baseline for researchers to identify more categories in the future.

While some conventions appear self-explanatory, without direct explanations from users we cannot know all their reasons for using these alternative strategies. Further, contacts within each scheme may be there for different reasons. Additionally, while it is clear that the intended use of the contact list does not match the actual usage in this dataset, we cannot make claims on what the shortcomings of existing contact lists are, and whether or not current designs address users' needs. These are compelling questions to explore in future work.

Finally, the design of services and applications a user chooses for managing her contacts (e.g. Google Contacts, Facebook, LinkedIn, Exchange, iCloud) influences how the user creates and maintains contact entries. The connections between those services, or whether they are joined at all, are a further influence. Future work should focus on the relationship between how contact entries are named and managed and the design of contact management services.

CONCLUSION

This analysis of contact lists from a broad range of 54 participants found that those lists were used in surprising ways and revealed consistent patterns. The behaviors we identified present both a challenge and an opportunity: though usage patterns prevent simple automated approaches for data mining or contact-list merging, they also suggest alternative directions for data mining to understand the behavior of individuals and their relationships with others.

ACKNOWLEDGEMENTS

This work was supported by Yahoo, The Stu Card Fellowship, A Google Faculty Research Award, National Science Foundation Grant No. DGE-0903659 and Defense Advanced Research Projects Agency program DCAPS Contract No. N66001-12-C-4196

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of the Defense Advanced Research Projects Agency of the U.S. Government.

REFERENCES

1. Bentley, F.R., Kames, J., Ahmed, R., Zivin, R.S., and Schwendimann, L. Contacts 3.0. *In Proc. CHI EA '10*.
2. Ekler, P. and Lukovszki, T. Experiences with Phonebook-Centric Social Networks. *In Proc. IEEE CCNC '10*.
3. Farnham, S.D. and Churchill, E.F. Faceted identity, faceted lives. *In Proc. CSCW '11*, .
4. Jones, W. *Keeping Found Things Found: The Study and Practice of Personal Information Management: The Study and Practice of Personal Information Management*. Morgan Kaufmann Publishers Inc., 2007.
5. Jones, W. *Keeping Found Things Found: The Study and Practice of Personal Information Management: The Study and Practice of Personal Information Management*. Morgan Kaufmann Publishers Inc., 2007.
6. Min, J.-K., Wiese, J., Hong, J.I., and Zimmerman, J. Mining smartphone data to classify life-facets of social relationships. *In Proc. CSCW '13*.
7. Whittaker, S., Jones, Q., and Terveen, L. Contact management. *In Proc. CSCW '02*.
8. Wiese, J., Kelley, P.G., Cranor, L.F., Dabbish, L., Hong, J.I., and Zimmerman, J. Are you close with me? are you nearby? *In Proc. UbiComp '11*.