P2P Decentralized Population Census

Song Qin[1] and **Marius C. Silaghi**[1]
Toshihiro Matsui[2] and **Makoto Yokoo**[3] and **Katsutoshi Hirayama**[4]
[1]Florida Tech, [2]Nagoya Inst. of Technology, [3]Kyushu Univ, [4]Kobe Marine Univ.

Abstract

We describe a framework and investigate techniques for running decentralized census processes that enable observers to independently verify governmental data. Census is a process impacting important issues such as the amount of funding that a community will get from a central government and its representation in the Congress. Correct census is essential for detecting vote stuffing. Census has been historically run by governments, but citizens and NGOs need to be able to verify it. Classical census is expensive and beyond the reach of these players, hence the need for affordable citizen-driven census technology.

Various citizens have different opinions as to what information should be gathered and what makes a person eligible to be accounted for in statistics. Using as inputs the official preferences of a given government should enable the verification of the data of that government. The reported work formalizes this problem and introduces a framework and techniques for running a fully decentralized citizen-driven census based on peer-to-peer technology. An addressed challenge is to quantify the uncertainty and the trust in the data provided by P2P users, as well as to develop techniques for reducing this uncertainty. We report on techniques to reason and to extract census-related conclusions based on the available data. Probabilistic models with various approximations are experimented for evaluating the census results in this context.

Introduction

We address the problem of gathering census data using a decentralized, citizen-driven mechanism. The challenge addressed here consists in formalizing the census problem and developing algorithms applicable in a peer-to-peer (P2P) approach.

Census processes have been run for thousands of years by governments as ways of estimating expected taxes and/or military power. Classically a census employs humans to talk to each resident, counting the population of a country as well as gathering certain data items about each individual, to be used in designing and justifying policy making. These processes have been run by governments and its power to alter the published figure has been identified as the main threat to stability in certain societies, as they enable large scale ballot stuffing (Analytica 2009). It is therefore in the interest of citizens and non-governmental organizations (NGOs) to verify and validate the results of the official census. For the sake of stability, it is also important for governments to increase the confidence of the citizens in official operations.

Running an independent census on a large scale and using classical procedures is a complex operation and most of the interested individuals cannot afford it. Achieving census results with the quality with which governments achieve them is unlikely in the absence of similar funding. However, it can be valuable and satisfying for some observers to be able to even partly verify and corroborate official data.

Example 1 *For example, assume an activist claims that the government miscounted his area reporting 6000 people instead of his estimation of 20000. If an independent census process trusted by this activist cannot return more than 6000 people, then the fears of the activist can be alleviated.*

Example 2 *Similarly, an activist may claim that a government has moved 20000 people in an area to change its ethnic composition. If an independent census cannot find more than a couple hundreds recent immigrants into that region, then fears may be alleviated. Alternatively, if a census identifies tens of thousands of recent immigrants, then claims can be corroborated.*

The results of an independent census may complement results of an official census (as it may give new opportunities for reaching additional residents. As such, governments can improve their official data using results from independent census processes, potentially after additional verification.

Eligibility The counting of residents as part of a census is done differently function of the philosophical principles of the given government (Earnest 2003a; Owen 2009; Beckman 2006).

Example 3 *In Switzerland, only citizens can vote at the federal level. However, in certain cantons (states)*

and municipalities, voting rights are granted to foreigners having lived there for some time, e.g., for ten years in Lausanne (Earnest 2003b).

Should a person residing in the area for the last 4 month be counted in the census? Since eligibility varies for regions, politics and even personal beliefs, the approach in this paper is to allow different semantics of eligibility to co-exist in the data gathering process. Each end-user of the census process can compute the final statistics according to her own principles and preferences, or according to the official set of preferences (when she is testing official results).

The domain of eligibility is captured by organizations. The concept of organization is comprehensive and is formalized in the later sections. As an intuitive illustration, an organization is defined by a statute that governs the way in which constituents are defined and the way in which they take decisions. Organizations can range from a club or a company, to a country.

For the management of census in large organizations we employ the concept of neighborhood. In our study neighborhoods are hierarchical, with the top of the hierarchy representing the global body of constituents. The lowest level is selected such as to define groups of constituents, preferably the largest such that each constituent in each group can verify with reasonable effort an identity claiming to belong to that same group (what is reasonable effort may depend on the organization and could, for example, be evaluated as one day of work). Another assumptions about a neighborhood is that everybody can verify the existence of neighborhoods that are siblings in the hierarchical tree to any ancestor neighborhood of the group to which it belongs (e.g., constituents in a city can verify with reasonable effort the existence of another given locality in their county, or of another state in their country).

Constituents of these organizations verify their census continuously by witnessing on each other's eligibility in the organization. Witnessing can be seen as an open vote, but a vote based on the reputation of the voter. While witnessing for a false identity can decrease the reputation of the voter, identities witnessed by a trusted source inherit some of that reputation.

Each observer trusts itself, and this defines a rooted graph of trust on which inferences relevant to the observer can be made. We experiment here with probabilistic models and other approximate methods of inference in these graphs of trust. The results from a set of volunteers is used to distinguish among techniques based on their usability and readability. Results based on simulated users and attackers are also investigated.

After introducing related work, in section Concepts we introduce the main definitions. Section Techniques introduces the experimented algorithms for evaluating user data. We conclude after discussing experimental results.

Background

Online White Pages directories (ATT Interactive Inc 2010) and online voter lists (Complex Systems Inc 2010), come the closest to the task of enumerating citizens, but none of them attempts specifically to be exhaustive. People can opt to be left out of White Pages. Also, White Pages list only people having a telephone line, and may list only a subset of the inhabitants linked to a given phone line. Voter lists are typically not available freely in their entirety (Complex Systems Inc 2010). They contain only users that voted in previous elections. Moreover, at one moment we noticed that one can edit the information of some voters without authorization and verification. Although in this case the information is no longer collected only by companies, but also by citizens, the correctness of citizen provided information cannot be verified.

One of the main challenges of large distributed collaborations is that one user can login under as many identities as she has time and desire to register. The creation and usage of such duplicated identities is referred in literature as the Sybil attack. The term Sybil attack was first introduced by (Douceur 2002) in a generic distributed computing environment. In the presence of a trusted authority, the resistance to Sybil attacks is either offered by explicitly certified participation as in Microsoft's Farsite (Thawte 2009) or by an implicit verification. This implicit verification can be regarded as too dependent on unsafe assumptions about underlying systems, as in the Cooperative File System (Dabek *et al.* 2001) which is a peer-to-peer storage system.

There is no globally trusted authority in our assumptions for the decentralized census process. That raises the problem of how to validate counterfeit identities. In (Douceur 2002), possible methods are classified into direct validation and indirect validation approaches. The former suggests that an entity only accepts identities that it has directly validated by some means. The latter suggests that an entity accepts identities that are vouched for by already accepted identities. Similar techniques are used in X509 certification schemes (Cooper *et al.* 2008). Our approach for validation of counterfeit identities is related to this, but we bring the idea of eliciting and exploiting both positive validations and negative validations for each identity.

Facebook launched the Social Login feature (Debjit 2011) to verify real users. Using this feature, users are given a few pictures of their friends and asked to name the persons in the pictures. This is more innovative than the approach using CAPTCHAS to verify real users. However, attackers knowing the person's friends can still name them in the pictures.

The concept of regional/neighbor based trust and verification is used in the Thawte Web of Trust (Thawte 2009). There, local trusted people called notaries can verify one's credentials and certify them using a Thawte certificate. Regional/neighbor based trust and verification is also used with PGP, where people meet for key signing parties, giving each other an independent proof

of identity after manually inspecting government issued documents.

Census processes with peer validation can be successful only if people are sufficiently connected to provide enough data to the decision making process. Studies of connectivity between people have been conducted in relation to existing social networks. A kind of constituency was discussed in (Fedoruk 2006).

Since the advent of the Internet users found a way to communicate and form groups. Initially we had the Bulletin Board Systems (BBS), then Usenet Groups but more recently we have witnessed a flood of online social network websites attracting millions of users. Most of the social networks are small-world networks (Milgram 1967). The small-world phenomenon was first introduced by Milgram (Milgram 1967) with his famous six-degrees of separation in social networks: everyone is six or fewer steps away from any other person in the world. More generally, a network is considered to be small-world if its average shortest-path value l grows logarithmically as a function of the number of nodes in the network.

A reputation system maintains scores inferred from other's opinions for participants. As discussed in (Resnick *et al.* 2000), many issues are still open in reputation systems. Notions of valued trust are proposed in (Yahalom, Klein, & Beth 1993) and extended in (Beth, Borcherding, & Klein 1994). The introduced values can be used to decide if an entity is sufficiently trustworthy. The values are inferred from a graph with nodes as entities and edges as the trust relations. They formalize trust relations of different types, among which are identification (ID) and trustworthiness (PR), and discuss the potential offered by networks of such relations to model known distributed authentication protocols. Their trust derivation algorithm is related to some of the approximated reasoning models in our work. One level fuzzy logic inference evaluation of participants (sellers and buyers) in reputation systems is encountered in (Song *et al.* 2005). For example, the reputation score of a seller on eBay can be inferred by goods quality or delivery time, etc.

Framework

In this section we introduce in detail the definitions of the items involved in a decentralized census. The concepts of peer, organization, and constituents used in this report are the ones we implemented in the `DirectDemocracyP2P` system.

Items used in this work are referenced using global identifiers (GIDs), built in a way that avoids intended and unintended collisions between different items (typically by generating them either as a public key, or as the secure digest of the items data). The secure digest function is denoted with $HASH(d)$ where d is the data whose digest is computed. Given a public key P, we refer to its secret key as $SK(P)$. The digital signature for data d using secret key S is computed by $SIGN(S, d)$.

Agents The software that works on behalf of the user to disseminate his actions and to gather and organize data generated by his peers is here referred to as a software agent.

Peer A peer refers to a user as represented by a set of software agents. Formally,

Definition 1 *A peer is specified by a tuple $\langle \mathcal{P}, p, a, s \rangle$ where $\mathcal{P}$ is the GID of the peer and is specified as its public key, p is the set of properties declared by the peer, such as name, email, etc.*

A signature s is computed as:

$$s = SIGN(SK(\mathcal{P}), p).$$

A set of Internet addresses a is also specified with the peer and is not signed, as it may contain any IP address ever used by this peer.

Grassroot Organization A grassroot organization is an entity representing a set of fixed rules that govern decision making for a certain group of people on a certain set of issues. This fix set of rules is also referred as *constitution*. The constitution specifies the mechanism used to define who is eligible to be a constituent and the mechanism to manage the constituency (e.g., information to provide at registration). Based on these rules, one can instantiate a virtual constituent assembly where the interactions between its constituents happen. For example, a grassroot organization within which a census of US inhabitants is computed would lay out the rules and mechanism of collaboration among volunteers interested in participating in the US census. The properties requested of its constituents at registration could include "Name", "E-mail", "Residential address", "Passport number", "Birth Location", etc.

For some organizations with very small constituency (university department, club), the census problem is trivial since everybody knows each other. A bottom-up census of shareholders for a large company (like IBM), may raise similar issues as the decentralized census of the inhabitants of a country or the members of a diaspora. Formally,

Definition 2 *A grassroot organization is a tuple $\langle \mathcal{O}, p \rangle$ where $\mathcal{O}$ is the GID of the organization, p is a set of parameters describing it and its constitution (or statute).*

Constituents The people with right to cast votes that have a predefined weight in an organization form its constituency. Not all members of the constituency are able to interact using software agents (due to availability, illness, age, or lack of skills). The members that do not control software agents and therefore do not generate items in the virtual space are referred to as *inactive constituents*. Users that directly generate items using software agents are referred to as *active constituents*, and typically they are a subset of the set of peers active in the system.

A constituent is defined by a tuple $\langle \mathcal{C}, \mathcal{C}', \mathcal{O}, i, d, r, s \rangle$ where $\mathcal{C}$ is its GID, $\mathcal{O}$ is the GID of the relevant organization, i is the set of identity details, and d is the date and time when i was declared. $\mathcal{C}'$ is the GID of the active constituent that submitted this information. For active constituents, $\mathcal{C}$ is specified as a public key, $\mathcal{C}' = \mathcal{C}$, r is the revocation status of $\mathcal{C}$, and the constituent data is signed with $s = SIGN(SK(\mathcal{C}'), \langle \mathcal{O}, i, r \rangle)$. With inactive constituents, $\mathcal{C} = HASH(\mathcal{O}, i)$, and the signature is $s = SIGN(SK(\mathcal{C}'), \langle \mathcal{O}, i, r \rangle)$.

Neighborhood For ease of accounting, constituents can be organized in tree structures with nodes (called neighborhoods) corresponding to localities, cities, counties, states and countries. In this case, localities, cities, counties, states and countries form a natural hierarchy. The leaf of the tree of neighborhoods is the smallest cell of the census management, and can be configured to correspond in real life to a block, a street or an area small enough (relatively to the population density) such that members can learn and easily verify residency of their neighbors.

Formally a neighborhood is a tuple $\langle \mathcal{N}, n, t, P, c, \mathcal{C}, \sigma \rangle$, where $\mathcal{N}$ is the GID of the neighborhood, $\mathcal{C}$ is the GID of a constituent supporting the existence of this neighborhood, n is the name of the neighborhood, t is its type/level (e.g., city, block, unit), P is the GID of the parent neighborhood ($\perp$ for a top neighborhood), and c is the list of expected types of descendant levels under this neighborhood. $N = HASH(n, P, c)$ $\sigma = SIGN(SK(C), \langle n, P, c \rangle)$

Witness Constituents in a grassroot organization can support or oppose the other constituent items' eligibility for being counted in a census. We say that they perform favorable or unfavorable witness stances for those identities. A witness stance can be associated with a set of semantic statements (as epistemological commitments associated to ontological commitments from a set Ω), such as:

- existence versus nonexistence of constituent name-address pair,
- active constituent public key (GID) belongs or not to the constituent with declared name-address pair,
- favored versus disfavored version of a multiply occurring constituent (e.g., at the current residence versus an old residence, or with a correct name versus a misspelled name),
- eligibility versus ineligibility of constituent,
- correctness versus inaccuracy of details in identity,
- reliability versus sloppiness of witness.

For example, when a constituent A declares constituent B to be a *sloppy witness*, then A believes that B does not carefully verify all the constituents that it witnesses, unlike a *reliable witness*.

Such a witnessing stance is defined by a tuple $\langle W, \mathcal{O}, S, T, m, e, d, \sigma \rangle$ where $\mathcal{O}$ is an organization identifier, S is the constituent identifier of the witnessing constituent, T is the constituent identifier of the target constituent item and e is an human readable explanation. The set of semantic statements of the witnessing, where each of them can be either *favorable* or *unfavorable*, is captured in m. The parameter d represents the creation time of this witnessing stance. The signature is generated as:

$$\sigma = SIGN(SK(S), \langle \mathcal{O}, T, m, e, d \rangle).$$

The GID of the witness stance is generated as:

$$W = HASH(\mathcal{O}, S, T, M).$$

Constituents can also witness about the legitimacy of a neighborhood. For example, they can state that no locality called *Geneva* exists in their county, or that no street called 21^{st} *Street* exists in their city. Such a witness stance is represented by a tuple $\langle W, \mathcal{O}, S, \mathcal{N}, m, e, d, \sigma \rangle$ where the only difference with witness stances for constituent items is that a GID of a neighborhood $\mathcal{N}$ is specified instead of the GID of a target constituent T. Semantic statements for such witness stances can be of type:

- favored versus disfavored version of a multiply occurring neighborhood (e.g., *New York* vs *New-York city*)
- existing vs nonexistent neighborhood

Census Process Concepts

Now let us introduce concepts involved in the decentralized census processes.

Citizen Interactions

A citizen-driven census requires participation of individual citizens for actions such as *residence declaration* and *witnessing*. As residence declarations, each individual voluntarily provides census data not only about herself but also about her neighbors. The neighborhood where a citizen resides is part of its identity details.

Verification A voting process, called witnessing, is used to help verify the census data. The verification can be done both by neighbors, and by volunteers who gather data about the inhabitants of the given area.

Witness Graph A graph defined by the witness relations between constituents can be generated in the following way:

- A node is generated for each constituent.
- A directed edge from node A to node B is generated for each semantic statement that A witnesses for B.
- Each edge has a color (from a set Ω), given by the type of statement that generated it (ontological commitment).

- An edge has weight 1 if generated for a favorable stance and weight 0 if generated for an unfavorable stance (epistemological commitment).

Inactive nodes are sinks for this graph. This graph can be used to reason about the eligibility of the declared identities and implicitly about the census.

Distributed Census Problem The Distributed Census Problem (DCP) for an observer Γ can be formalized as a tuple $\langle \mathcal{N}_\mathcal{S}, \mathcal{I}, \mathcal{P}_\mathcal{S}, \mathcal{R}, \mathcal{W}, \mathcal{M}_\mathcal{S}, \Gamma, \mathcal{O} \rangle$, where:

- $\mathcal{N}_\mathcal{S}$ is the set of neighborhoods $\mathcal{N}_\mathcal{S} = \{1, ..., d\}$,
- $\mathcal{I}$ is the set of person identities,
- $\mathcal{O}$ is the organization (constitution), defining the eligibility
- $\mathcal{P}_\mathcal{S}$ is a set of peers with identities from set $\mathcal{I}$, each peer $\mathcal{P}$ having a different interpretation of $\mathcal{O}$: $\mathcal{P}(\mathcal{O})$
- $\mathcal{R}$ is the set of residence declarations (constituent items)
- $\mathcal{W}$ is the set of witness stances
- $\mathcal{M}_\mathcal{S}$ is a model of the relation between the ground truth I^* and $\mathcal{N}_\mathcal{S}$, $\mathcal{I}$, $\mathcal{P}$, $\mathcal{R}$ and $\mathcal{W}$, as believed by the observer Γ (e.g., a certain belief network)

$\mathcal{I}^*$ (the ground truth), each having an identity from the set $\mathcal{I}$. The problem is to approximate the $\mathcal{I}^*$ that best explains $\mathcal{N}_\mathcal{S}$, $\mathcal{I}$, $\mathcal{P}$, $\mathcal{R}$ and $\mathcal{W}$ based on the model $\mathcal{M}_\mathcal{S}$.

Techniques

Here we present the techniques used to address the challenge of inferring a count of the constituency given a witness graph.

Eligibility Although anyone can participate in the census process of a grassroot organization, not everyone is eligible to be counted in the census. In a grassroot organization, which is the context of this study, the definition of eligibility is a function of the constituent. When the eligibility for a constituent is based on a subjective view, the census result is relevant only to the user (or users) sharing this view. Hence, we define the eligibility as a probabilistic function of several parameters:

- Someone's interpretation of the witness graph, $\mathcal{M}_\mathcal{S}$
- Someone's own definition of the eligibility, $\Gamma(\mathcal{O})$

Definition 3 *The reference user is the user Γ who currently computes the census.*

Definition 4 (Censable and Ψ) *A constituent item C is* `censable` *for an organization if it is eligible and new (never counted elsewhere). The Γ's confidence value in whether C is* `censable` *is denoted $\Psi(C)$.*

Definition 5 (Witness Reliability and Φ) *A constituent item C is a* reliable witness *if Γ trusts all the witness stances that C issues as she trusts her own. Γ may not fully trust the stances of another constituent C, but only with a confidence value $\Phi(C)$.*

Based on the DCP parameters, one can infer a value Ψ for the confidence that observer Γ can have on whether a given constituent item C identifies a `censable` user, and a value Φ for its confidence on whether C is *witness reliable*.

Remark 1 (Decision Criteria 1) *One approach to compute a census is to declare that an identity is eligible (to be counted in the census) from the point of view of the reference user Γ if the value of Ψ surpasses a threshold t where t is defined by the Γ.*

Remark 2 (Decision Criteria 2) *Another approach is to sum the values Ψ for all constituents (once normalized in the interval [0,1]).*

In the next paragraphs, we proposed algorithms to compute the Ψ value for each node in the witness graph for various types of models $\mathcal{M}_\mathcal{S}$.

For a given type of semantic statement, we will use the following concepts:

- The *supporting parents* of the node C that witness C favorably for the quality q, $q \in \Omega$, are denoted as $SP^q(C)$.
- The *opposing parents* of the node C that witness C unfavorably for the quality q, $q \in \Omega$, are denoted as $OP^q(C)$.
- The *supported children* of the node C are the children that are witnessed favorably by C for the quality q, $q \in \Omega$, denoted as $SC^q(C)$.
- The *opposed children* of the node C are the children that are witnessed unfavorably by C for the quality q, $q \in \Omega$, denoted as $OC^q(C)$.
- The *amortization factor* $f_q, f_q \in [0, 1]$, models the decrease of the confidence during transfer by witnessing for quality q. These factors compose as one gets further from Γ in the transitive chain of trust (along the *"reliable witness"* edges in the witness graph).

We introduce the notation Φ to denote the quality *reliable witness* when used as superscript or subscript with one of the notations above (e.g., $SC^\Phi(C)$). Similarly we use Ψ to denote the `censable` quality when used as superscript or subscript in these notations (e.g., $SC^\Psi(C)$).

Example 4 *Given a supporting parent node sp of the node C, the confidence $i_{sp,C}$ propagated (in certain introduced models) from sp to C is $\Phi(sp) \times f_\Phi$.*

Assume that a supporting parent sp of node C has value $\Phi(sp)$=0.8 and Γ's f_Φ is 0.9, the confidence $i_{sp,C}$ transferred from sp to C is 0.8×0.9=0.72.

Remark 3 *However, amortization may not apply to opposing parents even as it applies for supporting parents. In certain introduced models, given an opposing*

parent node op of the node C, the confidence $i_{op,C}$ propagated from op to C is 0.

The node representing the reference user Γ and its directly connected children are treated separately and referred to as *special nodes*.

Several of the proposed approaches share the following assumptions for the value of the *special nodes*:

Assumption `[Self_Trust]`

- $\Psi(a) = 1, \forall a \in SC^{\Psi}(\Gamma)$
- $\Psi(a) = 0, \forall a \in OC^{\Psi}(\Gamma)$
- $\Phi(a) = 1, \forall a \in SC^{\Phi}(\Gamma)$
- $\Phi(a) = 0, \forall a \in OC^{\Phi}(\Gamma)$

Approximate non-probabilistic Models

The models of approximate reasoning introduced next are: Max Amortized Support (MAXAS), Adjusted Max Amortized Support (AMAS), Average Support (AS), Penalized Average Support (PAS), and Adjusted Support Ratio (ASR).

Maximum Amortized Support (MAXAS) The first model we introduce for computing the Ψ value of each node in the witness graph is given in Algorithm 1. This model employs the amortization factor (Line 11) as per Example 4. The values of f_Φ and f_Ψ are user provided inputs to the algorithm. For each node, the initial Ψ and Φ are 0 since Γ a priori knows nothing about it (Line 1).

The *reference user*, Γ, who computes the census has full confidence in her witness stances (Lines 2 and 3), as per the assumption `Self_Trust`. Further, the algorithm traverses remaining nodes in the graph in breadth first order. If a node C has N *supporting parents*, $SP^{\Phi}(C)$, MAXAS computes $\Psi(C)$ as the maximum out of all the confidence values transferred from them (Line 11).

An example is given in Figure 1, and we use it next to illustrate the output of Algorithm 1. In this figure, for simplicity, a single edge is used to represent all semantic statements in a witness stance, and all of them are supposed to have the same weight (epistemological commitment, 1 or 0). The amortization factors are also equal: $f_\Phi = f_\Psi = 0.9$. Initially, each node's Ψ value is set to 0 (Line 1). Node S is the *reference user* Γ and the shown graph captures her data about other nodes (A, B, C, etc). The nodes are labeled with the confidence in their *witness reliability* that S infers from this graph. One can see that she trusts herself ($\Phi(S) = 1$, Line 2), and does not considers itself `censable`, ($\Psi(S) = 0$). S also trusts those for whom she issues stances as favorable *reliable witness* (A, B and D, $\Phi(A) = \Psi(A)$=1, $\Phi(B) = \Psi(B)$=1, $\Phi(D) = \Psi(D)$=1, Line 2). A, B and D are added to queue Q in this order. For node C, since $\Phi(C)$=0 we stop the distribution of trust to C's children, hence do not infer anything about F ($\Psi(F)$ and $\Phi(F)$ remain 0). A is then dequeued from the head

Input: A witness graph g with the starting node Γ and amortization factor f_Φ for the *reliable witness* quality and f_Ψ for the `censable` quality

```
1  for each node n in g do Ψ(n) ← 0; Φ(n) ← 0;
2  Ψ(a) ← 1, ∀a ∈ SC^Ψ(Γ); Ψ(a) ← 0, ∀a ∈ OC^Ψ(Γ);
3  Φ(a) ← 1, ∀a ∈ SC^Φ(Γ); Φ(a) ← 0, ∀a ∈ OC^Φ(Γ);
4  Add SC^Φ(Γ) to queue Q;
5  while Q is not empty do
6  |  node n ← extract_first(Q);
7  |  foreach c ∈ SC^Ψ(n) do
8  |  |  Ψ(c) ← max(Ψ(c), f_Ψ × Φ(n))
9  |  end
10 |  foreach c ∈ SC^Φ(n) do
11 |  |  Φ(c) ← max(Φ(c), f_Φ × Φ(n));
12 |  |  if c has never been added to Q then
13 |  |  |  add c to the end of the Q
14 |  |  end
15 |  end
16 end
```

Algorithm 1: Derivation of $\Psi(C)$ and $\Phi(C)$ for each constituent C

of the queue (Line 6). Since A has a favorable witness on D, now $\Phi(D)$ is evaluated to be $max(1, 0.9)$=1 (Line 11). We do not add D to the queue, since it is already in it. Then B is dequeued, but it has no favorable witness on its only child E. $\Psi(E)$ is not changed. Then D is dequeued. Since D has a favorable witness on E, $\Psi(E) = \Phi(E)$=$max(0.9, 0)$=0.9 (Line 11) and E is added to Q. C and F are never added to the queue.

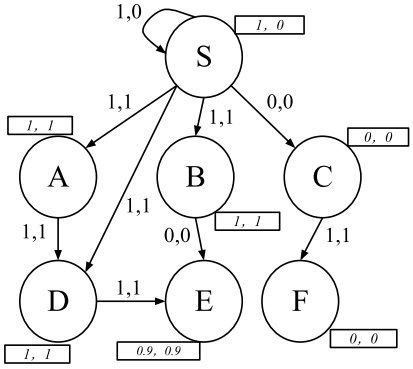


Figure 1: Propagation of (Φ, Ψ) value pairs

In the context of Algorithm 1, and when the census is estimated with the mechanism in Remark 1, in order to be counted, a node has to be within a certain support distance from the root s. The distance is determined by the used amortization factor and threshold. The smaller the factor, the smaller the required distance. In the example given by Figure 1, assume we choose the census threshold t as 0.95, S, A, B and D are counted. Note that E is not counted because its support path

exceeds the distance of 1=1 + $\lfloor \log_f t \rfloor$ edges from S.

Adjusted Max Amortized Support (AMAS) In this second model, to enable the increase of Ψ for a constituent when it gets extra support, we let $\Psi(C)$ take values between $M(C)$ and $N(C)$ ($\Psi(C) \in [M(C), N(C)]$) where:

$$M(C) = \max_{n \in SP^{\Psi}(C)} \Phi(n) \times f_{\Psi}$$

$$N(C) = \max_{n \in SP^{\Psi}(C)} \Phi(n)$$

The algorithm we use to compute the Φ value here is similar to Algorithm 1. We do not repeat the algorithm and only address the difference. In Algorithm 1, the Ψ value of a constituent item C is computed as $M(C)$. This model assumes Equation 1:

$$\Psi(C) = M(C) + \frac{(N(C) - M(C)) \times min(fw, W)}{W} \quad (1)$$

In Equation 1, fw is the total number of favorable `censable` witnesses ($|SP^{\Psi}(C)|$) for C and W is a user-defined parameter. The closer W grows towards $|SP^{\Psi}(C)|$, the closer $\Psi(C)$ approaches to $M(C)$. If C has more than W favorable witnesses from its parents, $\Psi(C)$ becomes $M(C)$.

Average Support (AS) In the first two approaches, note that $\Psi(C)$ is only inferred from the *supporting parents*. For the *opposing parents*, the propagation is stopped, i.e., the inputs from the *opposing parents* to the children nodes are 0s. In this approach and the next one, we employ the inputs from the *opposing parents* into computing the Ψ value of a node in the witness graph. AS can be seen as an extension with amortization factors of an interpretation of the method to combine recommendation trust proposed in (Beth, Borcherding, & Klein 1994), where recommendation values are replaced by the confidence in the witness reliability of constituents witnessing an entity.

The AS model computes the Ψ value using Equation 2 where $|SP^{\Psi}(C)|$ is the number of favorable witnesses for C and $|OP^{\Psi}(C)|$ is the number of unfavorable witnesses for C.

$$\Psi(C) = \frac{f_{\Psi} \times \max\left(0, \sum_{n \in SP^{\Psi}(C)} \Phi(n) - \sum_{n \in OP^{\Psi}(C)} \Phi(n)\right)}{|SP^{\Psi}(C)| + |OP^{\Psi}(C)|} \quad (2)$$

The trust associated with a constituent item is also computed using a similar expression:

$$\Phi(C) = \frac{f_{\Phi} \times \max\left(0, \sum_{n \in SP^{\Phi}(C)} \Phi(n) - \sum_{n \in OP^{\Phi}(C)} \Phi(n)\right)}{|SP^{\Phi}(C)| + |OP^{\Phi}(C)|} \quad (3)$$

Constituent	$\Psi(C)$	Counted Constituent
C_1	2.75	*Passed*
C_2	2.0	
C_3	2.0	
C_4	4.0	*Passed*
...	...	...

Table 1: Sample outcome with model ASR for S_w=6, O_w=4, t=2

Penalized Average Support (PAS) In the fourth discussed model, we build on the AS model but reduce the penalty introduced by opposing parents to only the part in the denominator of the fraction.

$$\Psi(c) = \frac{f_{\Psi} \sum_{n \in SP^{\Psi}(C)} \Phi(n)}{|SP^{\Psi}(C)| + 1 + f_{\Psi} \sum_{n \in OP^{\Psi}(C)} \Phi(n)} \quad (4)$$

The Φ value is also computed with a similar expression, just using the SP^{Φ}, OP^{Φ} and f_{Φ}, instead of the corresponding values for Ψ in Equation 4.

Adjusted Support Ratio (ASR) This last non-probabilistic approximate model is much simpler than previous ones. The heuristic used here is that the identity of a constituent is more likely to be considered `censable` by others if she has relatively more SP^{Ψ} than OP^{Ψ}. An equation reflecting this value is:

$$\Psi(C) = \frac{|SP^{\Psi}(C)|}{|OP^{\Psi}(C)|} \quad (5)$$

As computed in Equation 5, $\Psi(C)$ is respecting the heuristic in the sense that a larger SP^{Ψ} indicates a larger $\Psi(C)$ and well as a smaller OP^{Ψ}. However, division by zero has to be avoided, and the obtained range of $\Psi(C)$ may need to be adjusted for a given data. Hence S_w and O_w are added as user specified parameter to allow users to adjust the range of $\Psi(C)$. A sample output of the census process using Decision Criteria 1 is shown in Table 1.

$$\Psi(C) = \frac{|SP^{\Psi}(C)| + S_w}{|OP^{\Psi}(C)| + O_w} \quad (6)$$

Probabilistic Models

We will now discuss models based on Bayesian Networks. In this model, random variables are used to represent the `censable` property of each constituent, the *reliable witness* property, and the witnessing stances between each pair of constituents for each quality. All these random variables are Boolean. For each pair of constituents A and B we get the random variables and Bayesian Network in Figure 2. Note that each pair of constituents requires the introduction of $2|\Omega|$ random variables for $|\Omega|$ qualities. With the two considered qualities in Figure 2, Φ and Ψ, a constituent

CS^A	RW^B	$P(W^{BA\Psi})$
t	t	0.9
t	f	0.5
f	t	0.1
f	f	0.5

RW^A	RW^B	$P(W^{BA\Phi})$
t	t	0.9
t	f	0.5
f	t	0.3
f	f	0.5

CS^A	RW^A	$P(W^{AA\Psi})$
t	t	0.99
t	f	0.5
f	t	0.1
f	f	0.5

RW^A	$P(W^{AA\Phi})$
t	0.99
f	0.5

$P(RW)$	$P(CS)$
0.5	0.5

Table 2: Transition, sensor and prior CPT

(e.g., A) is associated with two hidden random variables: `censable` (CS^A) and *reliable_witness* (RW^A). Each pair of constituents items (e.g., A and B) is associated with four evidence (grayed) random variables: A witnesses for B being a reliable witness ($W^{AB\Phi}$), A witnesses for B being `censable` ($W^{AB\Psi}$), B witnesses for A being a reliable witness ($W^{BA\Phi}$), B witnesses for A being `censable` ($W^{BA\Psi}$).

While conditional probability tables can be trained from real data once large amount of such data is available, sample conditional probability tables built manually for variables of type $W^{BA\Phi}$ and $W^{BA\Psi}$ are shown in Table 2.

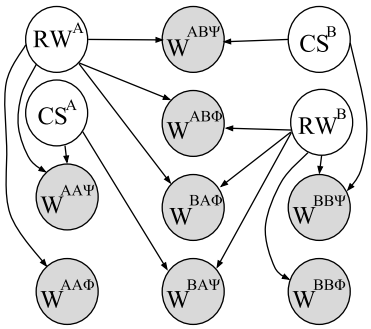


Figure 2: A Bayesian Network that models the DCP with two constituent items

Theorem 1 *The number of random variables (all Boolean) in a Bayesian Network modeling a DCP is linear in the size of the input.*

Proof Given n constituent items $C_1, ..., C_n$, we get at most $n^2|\Omega|$ random variables:

- $n|\Omega|$ hidden variables modeling the real qualities of each constituent item, and
- $n(n-1)|\Omega|$ modeling the evidence variables about all $|\Omega|$ possible witness stances between each of the $n(n-1)$ possible directed pairs of constituents.

Note that we do not need to model with random variable the nonexistent witness stances. Therefore the actual network size is linear in the size of the input, being proportional to the number w of input semantic statements in witness stances $(w+n)|\Omega|$. □

For average sized networks one can perform queries of values for the random variables CS^{C_i}, modeling $\Psi(C_i)$ of i^{th} constituent item, using techniques such as Markov Chain Monte Carlo (MCMC).

Experiments

To evaluate the power of the studied DCP models to represent users reasoning about census, as well as to resist various attacks, we perform two sets of preliminary experiments. One of them is based on a set of volunteers and the second is based on a larger simulated data.

In the experiments based on volunteers we asked 10 people living within an area of a few square kilometers to register themselves as active constituents and to also register others 10 friends as inactive constituents of a regional organization. Each of these volunteers had the opportunity to witness for the other constituent items that they knew. We also introduced 2 obviously wrong constituents at an address that most participants knew to not exist. A snapshot of the interactions between constituents is shown in Figure 3 where the thick edges represent favorable witness stances, the thin edges represents unfavorable witness stances, the nodes represents constituents and size of node is proportional to the in-degree. Red nodes denote active constituents and blue nodes denote inactive ones. Like in the example in Figure 1 of section Techniques, a single edge is used to denote all semantic statements in a witness stance.

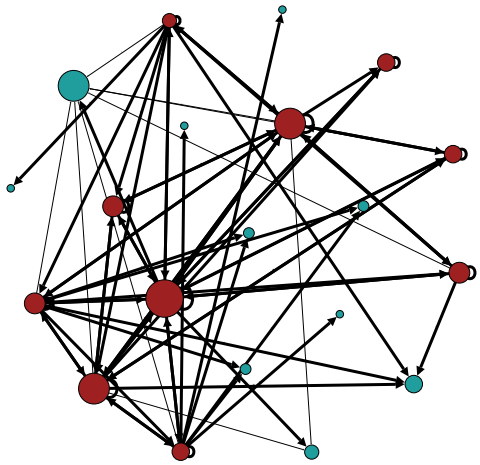


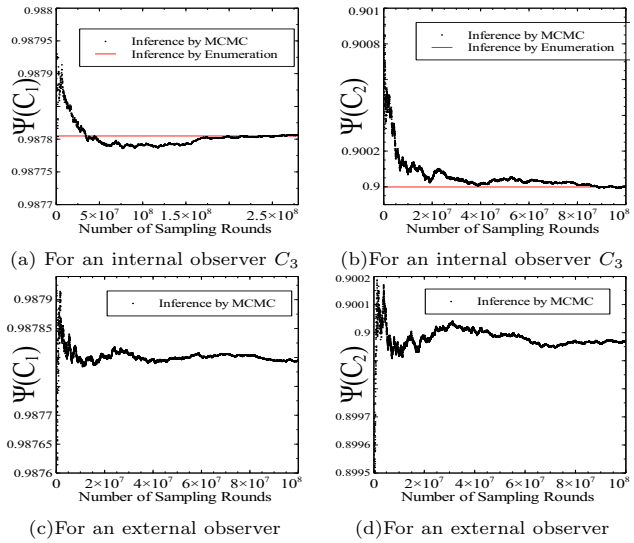
Figure 3: Visualization of constituents and the witness relation between them

After the P2P witnessing process reached quiescence, we asked 5 of the participants (who were available) to use the widgets implementing each of the available five non-probabilistic DCP models for deciding on a cen-

sus based on the available constituent items. Each of these constituents ranked the five models in terms of how well they were able to capture their own opinion on the `censable` status of each item and on their correctness. A score between 0 and 10 was assigned to each model. These scores are detailed in the Table 3. While the size of the sample is small and deviation of these scores is high, currently the winner is the Penalized Average Support model (PAS). It is remarkable that the ASR, which is a very simple computation performed also acceptably well.

Constituent	C_1	C_2	C_3	C_4	C_5	Average
MAXAS	8.5	5	10	3	7	6.7
AMAS	9.5	3	10	5	7	6.9
AS	8.5	5	10	6.5	7	7.4
PAS	9	8	10	6	7	8
ASR	10	2	6	8	9	7

Table 3: Models and their Scores. The preferred parameters for MAX are $t = 0.5$, $f = 0.5$, for AS are $t = 0.65$, $f = 0.7$, for PAS are $t = 0.43$, $f = 0.82$, for ASR are $|S_w| = 6$, $|O_w| = 4$, and $t = 2$.



(a) For an internal observer C_3 (b)For an internal observer C_3

(c)For an external observer (d)For an external observer

Figure 4: Inference of $\Psi(C_1)$ and $\Psi(C_2)$

With the Bayesian network described in Figure 2 and CPT specified by Table 2, we use MCMC to perform queries of values for the random variable CS^{C_i} which models the $\Psi(C_i)$ of the i^{th} constituents. Convergence for a few constituents is shown in Figures 4 (a), (b), (c), (d). The reference (red line) is computed with an exact inference by enumeration. For an external observer the exact inference by enumeration is expected to take 33 days on a computer and is not shown here.

While in the first set of experiments we mainly attempt to find the models most relevant to humans in representing the process of the census based on witnessing, the second set of experiments is targeting the evaluation of the robustness to attacks that such a system can provide.

Unlike for the case of experiments based on volunteers where our preliminary samples are small, we have performed extensive experiments with simulated data. Most of them will not be described here for lack of space. The result of a set of 5 experiments estimating the impact of the percentage k of honest active constituents(HACs) in the global population on the `censable` properties of the constituents is shown in Figure 5. The true positive rate (TPR) gives the percentage of correctly counted constituents out of the total number of eligible constituents. The false positive rate (FPR) gives the percentage of wrongly counted constituents with respect to the total number of eligible constituents. A robust census process has a high TPR and a low FPR. The semantic statements for witness stances are only about the eligibility quality. In this experiment, a constituent item (n, a) is eligible if there is someone whose name is n and lives at address a. We simulate attackers that declare a number of ineligible constituent items and perform favorable witness stances for a percentage of h ineligible constituent items in their leaf neighborhoods. We assume that witness stances represent all semantic statements. The plotted points are for h of 100%, 93.75%, 87.5%, 75%, 50%, 25%, 12.5%, 6.25%, 0% respectively. The studied values of k are (0.9, 0.8, 0.7, 0.5 and 0.3) respectively.

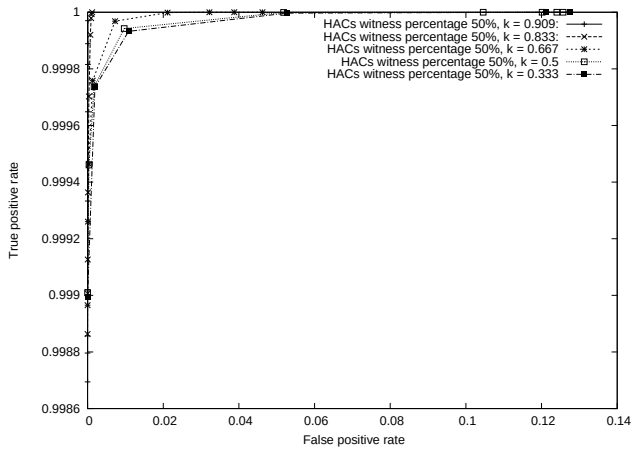


Figure 5: The effects of the percentage of honest active constituent (k value) in global population with curves defined by varying h.

There are several common parameters for the plotted curves. The total number of eligible constituents is 9300000. The *HACs witness percentage* is 50% out of their leaf neighborhoods. *HACs witness percentage* is the percentage of constituent items that an HAC witnesses honestly (A favorable witness stance is performed if an item is eligible and an unfavorable witness stance is performed if an item is ineligible). The num-

ber of *attackers* is 300000. The number of ineligible constituent items declared by each *attackers* is 4.

Since we see in Figure 5 that the curve with parameter k=0.9 is higher than the curve with parameter k= 0.8, the curve with parameter k=0.8 is higher than the curve with parameter k=0.7, the curve with parameter k=0.7 is higher than the curve with parameter k=0.5 and the curve with parameter k=0.5 is higher than the curve with parameter k=0.3, we conclude that the k value will affect the robustness of the system positively. That is, the bigger the k value is, the more accurate the system will be.

Conclusions

We have addressed the problem of formalizing and solving the decentralized population census problem (DCP). While population census is an important process with large implications in the distribution of public funds and security of elections from vote stuffing, it is currently an expensive process outside the reach of external verifiers and was identified as a threat to stability in certain regions.

To enable a decentralized citizen-driven population census, we investigate a set of concepts such as: agents, peers, grassroot organization, constituent, neighborhood and witnessing. Items for these concepts are identified by global identifiers guaranteed to be unique and that are disseminated among peers based on P2P protocols (current experiments being based on the DirectDemocracyP2P platform). A peer is an user acting under one name and public key via multiple agents (e.g., one agent per device that she uses).

The grassroot organization is a set of rules (constitution) the specify mechanisms to define eligibility of constituents. For large organizations, the constituency is organized in a tree of neighborhoods to help with census organization. Constituents can witness (vote) on each other's qualities, such as: eligibility and witnessing reliability.

A set of five efficient but approximate models of relations between witness stances and properties of constituents are proposed and empirically evaluated. We have also proposed and analyzed theoretically a probabilistic model based on Bayesian Networks that can be used to address the problem in a principal way. Preliminary experiments with volunteers are used to rank these models, while experiments with large simulated data show that robustness to attackers is possible when there exists a reasonable kernel of honest active constituents.

References

Analytica, O. 2009. Politics of census-taking threatens stability. `http://www.oxan.com/display.aspx?ItemID=DB154173`.

ATT Interactive Inc. 2010. `http://www.yellowpages.com/whitepages`.

Beckman, L. 2006. Citizenship and voting rights: Should resident aliens vote? *Citizenship studies* 10(2):153–165.

Beth, T.; Borcherding, M.; and Klein, B. 1994. Valuation of trust in open networks. *Computer Security-ESORICS 94* 1–18.

Complex Systems Inc. 2010. `http://www.floridavoterfile.com`.

Cooper, D.; Santesson, S.; Farrell, S.; Housley, R.; and Polk, W. 2008. Internet x.509 public key infrastructure certificate and certificate revocation list (crl) profile. *Network Working Group Request for Comments* 5280.

Dabek, F.; Kaashoek, M.; Karger, D.; Morris, R.; and Stoica, I. 2001. Wide-area cooperative storage with CFS. In *Proceedings of the eighteenth ACM symposium on Operating systems principles*, 202–215. ACM.

Debjit. 2011. Facebook launches social login and https to protect your privacy. `http://digitizor.com/2011/01/26/facebook-social-login-https/`.

Douceur, J. 2002. The sybil attack. *Peer-to-peer Systems* 251–260.

Earnest, D. 2003a. Noncitizen voting rights: A survey of an emerging democratic norm. In *2003 annual convention of the American Political Science Association, Philadelphia, August*, 28–31.

Earnest, D. C. 2003b. Voting rights for resident aliens: A comparison of 25 democracies. In *2003 Annual Meeting of the Northeast Political Science Association and the International Studies Association-Northeast.*

Fedoruk, A. ; Denzinger, J. 2006. A general framework for multi-agent search with individual and global goals: Stakeholder search. *International Transactions on Systems Science and Applications (ITSSA)* 1(4):357–362.

Milgram, S. 1967. The small world problem. *Psychology today* 2(1):60–67.

Owen, D. 2009. Resident aliens, non-resident citizens and voting rights: towards a pluralist theory of transnational political equality and modes of political belonging. In Calder, G.; Cole, P.; and Seglow, J., eds., *Citizenship Acquisition and National Belonging: Migration, Membership and the Liberal Democratic State.* Palgrave Macmillan. 52–73.

Resnick, P.; Kuwabara, K.; Zeckhauser, R.; and Friedman, E. 2000. Reputation systems. *Communications of the ACM* 43(12):45–48.

Song, S.; Hwang, K.; Zhou, R.; and Kwok, Y. 2005. Trusted p2p transactions with fuzzy reputation aggregation. *Internet Computing, IEEE* 9(6):24–34.

Thawte. 2009. Web of trust. `http://www.thawte.com/secure-email/web-of-trust-wot/`.

Yahalom, R.; Klein, B.; and Beth, T. 1993. Trust relationships in secure systems-a distributed authentication perspective. In *Research in Security and Privacy, 1993. Proceedings., 1993 IEEE Computer Society Symposium on*, 150–164. IEEE.