Using Trust to Resist Censorship in the Presence of Collusion

Andriy Panchenko and Lexi Pimenidis*

RWTH Aachen University,
Computer Science Department - Informatik IV,
Ahornstr. 55, D-52074 Aachen, Germany
{panchenko,lexi}@i4.informatik.rwth-aachen.de

Abstract. Censorship resistance deals with an attempt to prevent censors from the acquaintance and thus blocking of distribution of a particular content through the network. Providing resistance against censoring is a very challenging and difficult task to achieve. However it is vital for the purpose of freedom of speech, mind and achievement of democratic principles in todays society.
In this paper we define a model of a censorship resistant system. Thereafter we propose to split the problem of resisting censorship into the following two sub-problems: a trusted directory and steganographic data transfer. The directory is used in order to prolong contacts among peers based on their reputation in a way, that honest members get contacts only to other honest peers and colluded members remain isolated. Furthermore, we aim to provide an analysis of a trusted directory for reputation and its implications on censorship resistant systems. To this end we define a set of properties that such a directory has to fulfill and develop a proposal for the implementation. Finally we provide a simulation-based validation of our approach.

1 Introduction

According to the Universal Declaration of Human Rights, everyone has the right to freedom of opinion and expression, including receiving and imparting information and ideas through any media regardless of frontiers [9]. In todays world, however, an increasing number of organizations, companies and even countries block free access to parts of the Internet [10]. The censors try to impede accessing some special political, ethical or religious content. For example, Saudi Arabia runs a country-wide Internet Service Unit (all ISPs must, by law, route through it), which provides an infamous web-censoring system that is supposed to protect Saudi citizens from "those pages of an offensive or harmful nature to the society, and which violate the tenants of the Islamic religion or societal norms"[2]. Another well known example is the "Great Firewall of China", where

* The authors are funded by the European Commission's 6th Framework Program.
[2] http://www.newsforge.com/article.pl?sid=04/01/12/2147220

strict censoring is provided at the governmental level. Lots of web pages like the British radio station BBC, human rights organizations, or the free encyclopedia Wikipedia are blocked. According to an Amnesty International report there are 54 people in jail in China because of illegal content distribution[3]. International Internet search engines like Google, Yahoo and Microsoft's MSN were recently criticized for censoring search results according to China's guidelines. Moreover, content filtering is also a subject in democratic nations. So, for example, US Marines Corps censors web access for troops in Iraq[4,5]. The European Union considers filtering and ranking according to the Internet Action Plan [4].

For the purpose of freedom of speech, mind and achievement of democratic principles there is a great demand to withstand filtering and censoring of information access and dissemination. Blocking resistant[6] systems try to provide as much reachability and availability as possible, even to users in countries where the free flow of information is organizationally or physically restricted [6].

Censorship resistant systems often have to provide anonymity to its users in order to grant their protection (especially from the blocker) and therewith to achieve desired properties of the system. Providing resistance usually requires distributed, peer-to-peer systems in order to overcome the blocking of the central server entity. Distributing functionality across many network nodes allows to avoid an obvious single point of failure where an attacker can clog the entire network. Using peer-to-peer based systems, though, requires the need to place trust on peers in the network. For this purpose reputation can be introduced. However, if the main objective of the network is to provide support for anonymity, the realization of the reputation itself becomes very problematic. Hiding the real identity gives a possibility for an attacker to easily throw away a pseudonym that has acquired a bad reputation. Furthermore, it is difficult to verify a member's behavior while keeping his status anonymous as these are two contradictorary things. However, to the favor of blocking resistance, blocker and "normal" users have different objectives which can serve as an incentive for the classification.

2 Related Works

Zittrain and Edelman present their research results about Internet filtering practices by different countries and organizations worldwide in [10]. This includes country-specific results as well as studies of the concrete filtering software.

Perng et al. [7] define a term of *censorship susceptibility* (probability that an adversary can block a targeted document while allowing at least one other to be retrieved from the same server). Thereafter the authors analyze current implementation of censorship resistant schemes with respect to the defined model of

[3] March 2006, see also http://www.heise.de/newsticker/meldung/70800
[4] http://yro.slashdot.org/article.pl?sid=06/03/07/1613236
[5] http://wonkette.com/politics/wonkette/our-boys-need-gossip-158687.php
[6] We use terms "blocking resistance" and "censorship resistance" as synonyms.

censorship susceptibility. They call a system resistant to censoring if the censor must disable access to the entire host in order to filter the selected content. Further they show that existing systems fail to meet the above provided strong adversary definition. Authors claim that Private Information Retrieval (PIR) is necessary, though not sufficient to achieve the definition. Moreover, they propose to use PIR in combination with digital signatures in order to reach the required properties [7].

Danezis and Anderson [2] propose an economic model of censorship resistance inspired by economics and conflicts theory. They assess how two different design philosophies – random and discretionary (encouraging nodes to serve content they are interested in) distribution of content in peer-to-peer network – resist censorship regarding to the model. The main finding was that a discretionary distribution is better suited to solve the problem.

Köpsell and Hillig [6] have proposed mechanisms in order to extend blocking resistance properties of their anonymity service AN.ON. The proposed principles are not technically mature in the sense that they do not solve the entire problem, but rather can only be used to make the job of the blocker more difficult. The work gives, though, a very good overview of the problematic and possible directions that solutions should strive for.

Infranet [5] is a system developed at the MIT that enables surreptitiously retrieval of the censored information via cooperating distributed web servers. The system uses a tunnel protocol that provides a covert communication channel between clients and servers. The latter also provides normal uncensored content. The requests are hidden by associating meaning to the sequence of HTTP request, and the results are placed into uncensored images using steganographic techniques.

Some other examples of censorship resistant systems are Freenet [1], Free Haven [3], Publius [13], Tangler [12], etc. Generally it is possible to say, that all known systems try to establish as many entry nodes to the blocked network as possible [6]. The idea is to hope that the blocker is not able to block all those nodes.

3 Model

In this section we explain our view on censorship resistant systems and explain in detail the model and level of abstraction that we want to use in the following parts of the paper.

For the simplicity of explanation we call all regular users that are part of a censored system *Alices*, those on the side which is not subject to filtering – *Bobs*, and the guardian entity – *warden*. Let $\mathcal{A}$ be the set of Alices and there exists a subset $\mathcal{A}' \subseteq \mathcal{A}$ that cooperates with warden $\mathcal{W}$. Let $\mathcal{B}$ be the set of Bobs. There also exists a subset $\mathcal{B}' \subseteq \mathcal{B}$ that cooperates with warden $\mathcal{W}$. The adversary can thus be seen as $\mathcal{W}' = \mathcal{W} \cup \mathcal{A}' \cup \mathcal{B}'$. Finally, there exists a group of users $\mathcal{A}^* \subseteq \mathcal{A} - \mathcal{A}'$ that are interested in communication with entities from

the set $\mathcal{B}^* \subseteq \mathcal{B} - \mathcal{B}'$ on some specific topic that the warden wants to censor. Initially there exist neither $a \in \mathcal{A}$ nor $w \in \mathcal{W}'$ that knows which $b \in \mathcal{B}$ are also in $\mathcal{B}^*$.

All users $a \in \mathcal{A}$ and $b \in \mathcal{B}$ relay messages to each other through $\mathcal{W}$. These messages can be of two types:

- *"good"* - those that do not include information that the warden is interested to censor;
- *"bad"* - are those the warden wants to filter.

It should be mentioned, that the guardian only profits from filtering "bad" messages, since blocking "good" has negative impact on his utility (e.g. consider losses of the Chinese economy from blocking trading transactions).

Based on educated guesses regarding the type of the message and its sender and receiver, the warden can choose to do one of the following actions:

- forward original data;
- forward modified data;
- drop data.

Moreover, he may store messages in order to collect enough evidence about his suspicions regarding some Alices. Furthermore, based on his suspicions he may (possibly even physically) punish the sender. The risk of $a \in \mathcal{A}^*$ to send messages of type "bad" to $b \in \mathcal{B}$ rises with the probability of correct classification of messages by the warden (if the warden can correctly classify messages the risk of detectibility is 100%, if he cannot do the classification there is only a marginal risk).

We call a system that allows any $a \in \mathcal{A}^*$ to communicate with $b \in \mathcal{B}^*$ despite the existence of $\mathcal{W}'$ *'blocking resistant'* with respect to the properties of $\mathcal{W}'$. Note that the probability of messages of type "bad" being blocked is not necessarily linked to the probability of being correctly classified by the warden (and vice-versa: not being blocked is not a sign for not being detectable). This is due to the fact that messages can be stored and analyzed off-line by the guardian. Thus the communication can take place but the evidence remains.

4 Our Approach

One way to achieve blocking resistance is to split the problem into two parts which can be solved separately:

- finding $b \in \mathcal{B}^*$;
- communicating with $b \in \mathcal{B}^*$.

The latter can be achieved by means of steganography [6]. It is possible to use one of the following for the first part – finding $b \in \mathcal{B}^*$:

- extensive search - this is a basic discovery technique that can be applied in the address space of $\mathcal{B}$;

– secret channel - it is possible to assume that there exists some small bandwidth information flow from Bobs to Alices. This can be provided e.g. by means of a satellite broadcast channel or some other channel that is not controlled by the warden but whose costs are much more expensive. This information can be especially effectively used for bootstrapping a censorship resistant system;
– a collusion resistant probabilistic directory for answering contact queries.

Extensive search is a very time and resource consuming procedure. Moreover, it is very problematic to place trust on newly discovered nodes since these can be warden's agents. The secret channel approach has additionally scalability issues, i.e. consider the dissemination of information about the channel on both communication sides. A collusion resistant directory to find peers to communicate with seems to have obvious advantages against the other two techniques. However, it should be investigated whether it is possible to build a directory that protects the presence of honest users from the warden and its agents and at the same time prolongs contacts among system users.

It should be noted that the warden possibly has a huge amount of resources at his disposal, however, the human ones remain most expensive and cannot be raised as easily as e.g. computational power. For this reason it is possible to generate tests by $\mathcal{B}^*$ that most humans can pass but are infeasible for automatic programs (in order to tell humans and computers apart) as described in [11]. This could be used in order to distinguish regular users from the warden's automated attempts to get contacts.

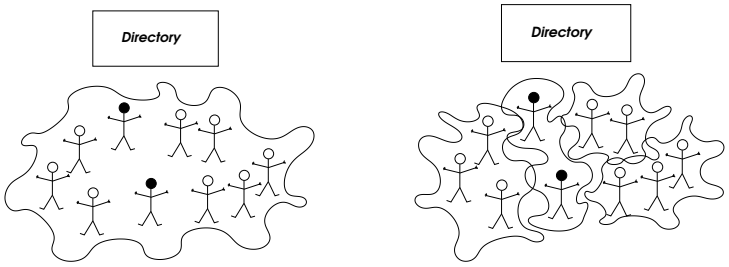
Blocking Resistant Technique

We propose to split the problem of creation of a censorship resistant network into the following two sub-problems:

– net of trust;
– steganography.

The net of trust can be implemented as a directory and can then be used by $a \in \mathcal{A}^*$ to find communication partners in $\mathcal{B}^*$. This puts some degree of trustworthiness on the contacts from directory: if it works as it should and can distinguish honest users from colluded ones, the contacts that are provided by the directory to the honest users are much more trustworthy than e.g. some random addresses. Steganographic communication is necessary to hide the traffic to and from the system, as well as between users.

Up from now we assume that all communication takes places in some steganographic form. This is necessary to thwart the threat that the warden can distinguish between ordinary messages and those belonging to the censorship resisting system. Given some sufficiently stealthy technique, the warden remains with only two ways to compromise the system: either he owns the directory or he tries to subvert the directory by inserting colluded members.



(a) Initial Setup (b) Situation After Some Epochs

Fig. 1. The Effect of Clustering by Directory

While we can exclude the first possibility as uninteresting (w.l.o.g. it is possible to assume the presence of the directory at the side of Bobs; digital signatures can be used in order to prevent impersonating), the latter remains a threat.

Collusion Resistant Directory: Properties

The directory should be distributed (in order to provide protection against denial-of-service attack and single point of failures) and to make it difficult for the warden to block the access to the directory. At first, however, we want to investigate the applicability of the approach with the help of a central directory. If the centralized directory cannot provide protection for the net of trust and is not resistant against the *"domino"* effect (having detected some user $u \in \mathcal{A}^* \cup \mathcal{B}^*$ the warden must learn as little information as possible from this fact about the other users in $\mathcal{A}^* \cup \mathcal{B}^*$ and their communication), it is most probable neither will be a distributed directory. One of the reasons for this is that the centralized directory has a global view on all members in the system and receives a feedback from all communication partners, thus gives agents less possibilities to have unnoticed dual behavior. Due to the fact that constructing a distributed directory requires much more effort, we make an analysis for the centralized one first. If it is not possible to build a centralized directory with the above mentioned properties, we suppose that neither would be a distributed one suitable to satisfy the requirements.

The directory will need to support the following functionalities in order to fulfill its duties:

Join Users must be able to establish new identities in this directory. There is no need for the directory to either limit the number of accounts per user, or to know the user's real identity. However, users must be in possession of a proof of the claiming identity if they want to reuse it.

Poll Once a user has an account on the system, he is allowed to query the directory for addresses of other users. To thwart the possibility that users

drain the system's resources, they receive only a single address per time interval, e.g. a day.

Push Users may return feedback to the directory, expressing to which extend they "trust" a peer. The directory should take care that users and their communication partners only submit feedback for peers for which they received the address from the directory. It must be possible to change the trust value at any point in time if users make new experiences.

After receiving a pseudonymous address of a contact from the directory the user makes an experience with the negotiated peer and both parties send feedback back to the directory. Based on this feedback the directory should distinguish users in the way that the probability of finding (getting contact to) $b \in B^*$ should be

- high for $a \in \mathcal{A}^*$;
- low for $w \in \mathcal{W}'$.

We claim that in order to be resistant against collusion it suffices to have steganographic communication and a directory with the properties as mentioned above. At this point, we will not go into steganography, but rather assume that there are systems that allow hidden communication like e.g. [5].

Collusion Resistant Directory: Methodology

In order to achieve collusion resistance, the directory must make sure that either no colluded user is able to harvest many of the honest users' addresses, or that the cost of doing so is prohibitive high such that it is not worth doing so.

To this end, we propose that the directory clusters its users into disjunct sets, where each user will only be able to receive addresses from other users within the same set. Therewith, if the directory manages to correctly classify different groups of users, e.g. honest Alices, Bobs and those cooperating with the warden, the honest users will be able to find each other, while the colluded members will only be referred to other colluded members.

One method of grouping its users into disjunct sets is to run a clustering algorithm using the users' trust vectors as an input. In the next section we will describe the procedure in detail and show the results of the evaluation that are produced by our technique (in order to test its suitability for the purpose of trusworthy contacts' dissemination). The desired result is depicted in Figure 1. As already mentioned, the warden's agents are depicted with black heads. In the beginning (a) all system participants belong to the same cluster. After some initialization epochs, the warden's agents are isolated (b) and only get contact to each other.

5 Evaluation and Analysis

In this section we will show the findings of using cluster algorithms for directories in order to achieve collusion resistance.

Before investing effort in design and deployment of a real system and gathering data in the real world, we wrote a simulation to check the properties and applicability of this approach. The simulation was written in Python using the library `pycluster` from the University of Tokyo [8].

In the simulation we had an overall number of U users of the directory. The users arrived equally distributed over the complete time interval of the simulation and consisted of $U_h < U$ honest users and $U_c = U - U_h$ colluded users. The directory clustered the users into k disjunct clusters using the $k-means$ clustering algorithm [8] based on the Euclidean distance of the users' trust vectors (how they are trusted by the others).

Social Model

All users had a fixed "social intelligence" factor $\mathcal{I}(user)$ that was used to calculate how well they were able to distinguish other users' intentions, as well as to hide their own identity. The values ranged from zero to ten, where ten was used for simulated persons that were able to pretty good understand other person's intentions after several rounds of interactions. We interpreted the value of five in the way that the user would be of average intelligence, while the value zero would denote complete sillies.

The "level of trust" between two users started out being neutral, i.e. five on a scale from zero to ten, where zero denotes absolute distrust and ten absolute trust. We denote it as $\mathcal{T}_i(u,p)$ and calculate the level of trust that an honest user u places on his communication partner p after the i-th round of interaction by:

$$\mathcal{T}_i(u,p) = \begin{cases} \mathcal{T}_{i-1} \cdot \lambda + (1-\lambda) \cdot \xi \cdot \theta_h : \text{if } p \text{ is honest} \\ \mathcal{T}_{i-1} \cdot \lambda + (1-\lambda) \cdot \xi \cdot \theta_c : \text{if } p \text{ is colluded} \end{cases} \tag{1}$$

where

$$\theta_h = (\mathcal{I}(p) + \mathcal{I}(u))/2, \tag{2}$$

$$\theta_c = (\mathcal{I}(p) - \mathcal{I}(u) + 10)/2, \tag{3}$$

and λ is a factor that determines the influence of the previous trust value (before an experience of the last interaction), ξ is a fuzziness factor. We used a random value within the range of $[0.8, 1.2]$ for ξ.

In contrast to these, colluded members have always applied the following formula in order to make the trust vectors of other colluded users similar to those of honest members:

$$\mathcal{T}_i(u,p) = \mathcal{T}_{i-1} \cdot \lambda + (1-\lambda) \cdot \xi \cdot \theta_h. \tag{4}$$

User notified the directory after each contact about the changes of their trust vector. This way, the directory could cluster users based on the way they were trusted by others. We define a trust vector of user j as:

$$\mathcal{T}^j = (\mathcal{T}(u_1, u_j), \ldots, \mathcal{T}(u_n, u_j)), \tag{5}$$

where $\mathcal{T}(u_j, u_j) = \mathcal{T}_{max}$.

Simulation Workflow

Each time interval the simulation first checked whether new users had to join the network and, if it is the case, initialized new entities. It then simulated the directory and clustered the users based on their trust vectors (how they are trusted by the others) and according to the main simulation parameters (like e.g. number of cluster). Each user then queried the directory for an address of another user (which possibly returns an address he already knew). Immediately afterwards the users returned their feedback, i.e. inserted a new trust value for their peers or refined the existing value based on the interaction experience according to the formulas shown above. After a fixed amount of time slices we collected the final results and those from intermediate slices. The simulation was repeated several times in order to get rid of random noise and to pinpoint the deviation of the results, which yielded to be very small.

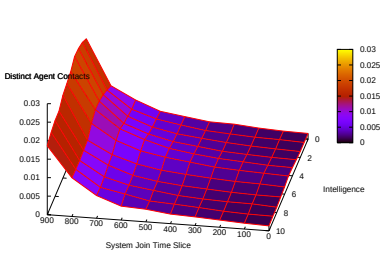
Results

The main results are shown in Figure 2. In that series of simulations we used $U_h = 100$ honest users, $U_c = 5$ on (a) and (b), $U_c = 7$ agents on (c) and ran each simulation over 1000 time slices. Users have joined the system equally distributed over the entire time interval.

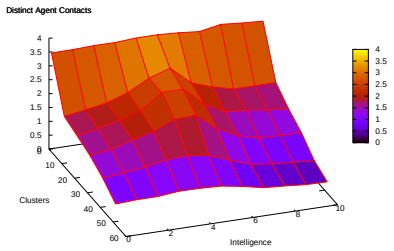
Part (a) shows on one axis the value of honest users' "social intelligence", while the second axis displays the time slice at which the users had joined the system. The z-axis displays the average number of distinct agent contacts for the honest users that joined the system at the corresponding time slice. Of great interest is the finding that the longer a user stays in the system, the less contacts to colluded members he gets with the time.

Part (b) of Figure 2 displays the effect of rising the number of clusters on the average number of distinct colluded users that each honest user communicated with. While two clusters were obviously not enough and nearly all users were seen by all five colluded members, the number dropped significantly for ten clusters and dropped even further for fifty. The small hill in the middle of picture (b) is due to the fact that the agents actively dislike honest users with a low intelligence, while honest users with a higher intelligence dislike agents.

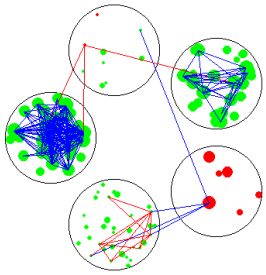
Picture (c) shows an example clustering at the end of a simulation run with 5 clusters: while nearly all agents (red dots) are in the same cluster, there are two main clusters each with roughly half of the more intelligent honest users (large green dots), and two clusters for those with less social intelligence (small green dots). The size of the circle reflects the intelligence of the agent. The blue lines denote trust between peers (trust level is not smaller than 8.5), the red one – distrust (trust level is not greater than 0.5).



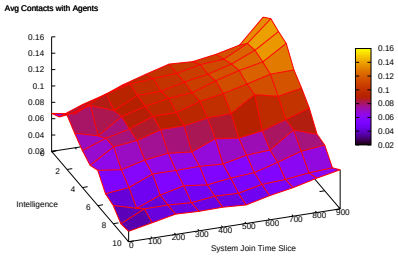
(a) Average Distinct Contacts With Agents



(b) Average Distinct Contacts With Agents



(c) Simulation Outcome: Trust Network and Clusters



(d) Average Contacts With Agents

Fig. 2. Results

Discussion

With the above provided results we have shown that clustering users based on their mutual trust seems to be a promising method for building a collusion resistant directory.

Besides clustering users on basis of the vector of how they were trusted, we also experimented with clustering users on basis of how they trust the others. While the results were similar, we finally chose the vector of how a user was trusted because this way it is much harder for colluded users to successfully attack the clustering algorithm. An evil node may try to guess the values that are assigned by the honest members. Moreover, colluded nodes may cooperate and place "catches" – having produced different values for the trust vectors they will be able to catch different classes of honest system users. Clustering based solely on the single value of (dis-)trust between two persons was not significantly better than choosing random contacts. Moreover, in order to change the default trust value in this case, users have to communicate with each other at least once, which is not desired.

There are two major different ways to attack this scheme: the first way is to poison the database with a lot of different entries. But as long as real

users act sufficiently different from automated entries, and as long as users do not really interact with the poisoned entries[7], this does not seem like an easy way to subvert the scheme. The seconds way is to convince other users that a colluded user is an honest one. In order to achieve this, an agent has to behave like an honest user over a long period of time, i.e. in interests of honest users. Therefore agents have to provide service that is not in their interest over a continuous period of time for "catching" dissidents with a high intelligence. Further research on the economics of playing double role for agents has to be performed. Also the impact of agents to honest users ratio on the system behavior has to be investigated.

6 Conclusion

In this work we have defined our model of a censorship resistant system and proposed to split the problem into a net of trust and steganographic data transfer. Steganographic communication is necessary to hide the traffic to and from the system as well as between the users. The net of trust is needed in order to find peers for communication and prolong contacts among them. We have proposed to realize it as a collusion resistant, probabilistic directory. A definition of a set of properties has been given that this directory must fulfill. With the simulation based evaluation we have shown that clustering users based on their trust vectors is a very promising method to build a directory with the defined before properties. With the help of the clustering algorithm, the trusted directory has become a powerful tool to distinct between different user classes *without* classifying them as *"good"* or *"bad"*. We achieve this by clustering the system users into disjoint sets, instead of calculating a global value of trustworthiness.

To ease the implementation we have investigated the approach of a centralized directory. In order to provide protection against denial-of-service attack, single point of failures and, not less important, to make it difficult for the warden to block the access to the central entity, switching to distributed directory and its implications must be researched and implemented.

All in all it is hard to say at this point to which extent our results are applicable to real systems. Even though we took care to choose powerful social model, it is very difficult to sufficiently abstract and simulate the human behavior and interpersonal trust. Therefore, in order to make final conclusion statements about our approach, evaluation in real world settings are necessary.

References

1. Ian Clarke, Oskar Sandberg, Brandon Wiley, and Theodore W. Hong. Freenet: A Distributed Anonymous Information Storage and Retrieval System. In *Proceedings of Designing Privacy Enhancing Technologies: Workshop on Design Issues*

[7] It is as difficult as solving the Turing Test to create a program that interacts as humans.

in Anonymity and Unobservability, pages 46–66, July 2000. `http://citeseer.nj.nec.com/clarke00freenet.html`.
2. George Danezis and Ross Anderson. The economics of censorship resistance. In *Proceedings of Workshop on Economics and Information Security (WEIS04)*, May 2004.
3. Roger Dingledine, Michael J. Freedman, and David Molnar. The Free Haven Project: Distributed Anonymous Storage Service. In H. Federrath, editor, *Proceedings of Designing Privacy Enhancing Technologies: Workshop on Design Issues in Anonymity and Unobservability*. Springer-Verlag, LNCS 2009, July 2000.
4. European Union Internet Action Plan: Filtering & Rating. http://europa.eu.int/information_society/activities/sip/projects/filtering/, 2006. visited March 2006.
5. Nick Feamster, Magdalena Balazinska, Greg Harfst, Hari Balakrishnan, and David Karger. Infranet: Circumventing web censorship and surveillance. In *Proceedings of the 11th USENIX Security Symposium*, August 2002.
6. Stefan Köpsell and Ulf Hillig. How to achieve blocking resistance for existing systems enabling anonymous web surfing. In *Proceedings of the Workshop on Privacy in the Electronic Society (WPES 2004)*, Washington, DC, USA, October 2004.
7. Ginger Perng, Michael K. Reiter, and Chenxi Wang. Censorship resistance revisited. In *Proceedings of Information Hiding Workshop (IH 2005)*, June 2005.
8. Pycluster - Clustering Library. http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm, 2005. visited March 2006.
9. Universal Declaration of Human Rights. http://www.un.org/Overview/rights.html, 1998. visited December 2007.
10. Documentation of Internet Filtering Worldwide. http://cyber.law.harvard.edu/filtering/, 2003. visited March 2006.
11. L. von Ahn, M. Blum, N. Hopper, and J. Langford. Captcha: Using hard ai problems for security. In *Proceedings of Eurocrypt*, pages 294–311, 2003.
12. Marc Waldman and David Mazières. Tangler: a censorship-resistant publishing system based on document entanglements. In *Proceedings of the 8th ACM Conference on Computer and Communications Security (CCS 2001)*, pages 126–135, November 2001.
13. Marc Waldman, Aviel Rubin, and Lorrie Cranor. Publius: A robust, tamper-evident, censorship-resistant and source-anonymous web publishing system. In *Proceedings of the 9th USENIX Security Symposium*, pages 59–72, August 2000.