A Position Paper on Data Sovereignty: The Importance of Geolocating Data in the Cloud

Zachary N. J. Peterson
Naval Postgraduate School

Mark Gondree
Naval Postgraduate School

Robert Beverly
Naval Postgraduate School

Abstract

In this paper we define the problem and scope of *data sovereignty* – the coupling of stored data authenticity and geographical location in the cloud. Establishing sovereignty is an especially important concern amid legal and policy constraints when data and resources are virtualized and widely distributed. We identify the key challenges that need to be solved to achieve an effective and un-cheatable solution as well as propose an initial technique for data sovereignty.

1 Introduction

The exponential growth of electronic data has led private organizations and governmental agencies with limited storage and IT resources to outsource data storage to cloud-based service providers. Storage service providers agree, by a service level agreement (SLA) contract, to preserve and make data available for retrieval for some level of durability. In addition to availability, many SLAs also guarantee that data will be stored only at data centers within a specific geographical region (*e.g.* within a state, time zone or political boundary) for performance, regulatory and continuity reasons. Actually *verifying* that cloud storage service providers are meeting their contractual geographic obligations, however, is a challenging problem, and one that has emerged as a critical issue. For example, careless or naïve storage service providers may move data, in violation of an SLA, to an overseas data center to leverage cheaper IT costs. Such actions, however, may make data available to foreign governments through search warrants or other legal mechanisms. A dishonest storage provider may intentionally move data overseas, to more easily leak information or to avoid legal liability.

In this position paper, we propose the need for developing new algorithms for establishing the integrity, authenticity, and geographical location of *data* stored in the cloud. Of particular interest is establishing data location at a granularity sufficient for placing it within the borders of a particular nation-state. We call this notion *data sovereignty*. We desire to establish some (probabilistic) guarantee that a provider is storing data at some expected physical location(s) and maintain such guarantees amid potentially dishonest providers. The problem of verifying that data exists *only* at allowed locations—and copies have not moved to some location that violates a policy—is a difficult problem in general; data sovereignty provides a much weaker guarantee, but a step toward actively monitoring compliance with some SLA policies.

Within the problem of data sovereignty, key concerns include developing techniques that minimize storage and network (thus, economic) costs. The immense size of digital archives, and the even larger size of the data centers on which they reside, make linear schemes (*i.e.* schemes that access every block of every file) prohibitively expensive, for either an auditor or storage provider. We posit that data sovereignty may not be solvable using any one technology, but rather may be achieved with a suite of existing and future tools that both detect and deter malicious behavior. Tools like these, which break the abstractions of the cloud to geolocate data, may be essential in the future to gather evidence, establish compliance (or show non-compliance) with contracts and laws.

2 Motivation

Most industries and governments are considering leveraging the scalability, cost, rapid provisioning and other benefits of the cloud. Those that are not, are at least considering the reality that—to enjoy new technologies, to stay competitive, or to remain relevant—they may be forced to follow this overwhelming technology trend and move business into the cloud. Moving to the cloud, however, requires organizations to interact with their data at a new level of abstraction. This comes with significant benefits, but also some limitations.

For example, a data owner may wish to store backups at multiple remote sites to provide resilience against natural disasters or for other continuity planning. Or, a data owner may desire her data be located physically near her target customers, for performance reasons. Rather than actively monitor QoS from the perspective of those customers, it might be more straight-forward to monitor the data's location. The cloud's abstractions, however, undermine the ability for a data owner to choose or control location, outside trusting a provider to meet some agreement. The US Federal Cloud Computing Strategy [12] outlines "actively monitoring service level agreements" and holding vendors accountable for failures as an essential part of any agency's cloud migration strategy. Data sovereignty protocols would provide such an active monitoring tool, for detecting compliance with SLA provisions concerning data geolocation.

Data sovereignty protocols may also be a complementary technology providing solutions to other data security problems. For digital provenance, when determining the origin and history of a digital document, one of the most fundamental questions is: where is this data, *right now*? With no reliable answer to this question at any point in the data's lifetime, one may never establish reliable provenance data.

Data sovereignty will also be beneficial to honest storage service providers. Even when data sovereignty protocols come at additional cost (*e.g.* performance or economic) we posit service providers may pay these costs to establish a level of trust with their clients, to comply with data retention legislation, or to meet new contractual obligations and remain competitive in the marketplace.

3 The Scope of Data Sovereignty

Data sovereignty has been recognized—although, not given a name—by cloud practitioners as a critical issue [15]; however, to date, the problem has been neither solved nor even posed in the research literature. As with any new security problem, it is important to clarify what a protocol should and should not attempt to accomplish. We propose that a proof of data sovereignty protocol should guarantee the possession, integrity and location of an instance of data in large networks that are not under an interrogator's control (*i.e.* the Internet). The data holder may act dishonestly (and possibly collude with other parties) to falsely claim to be holding a particular piece of data at some geographic location.

We presume the adversary to any data sovereignty solution is stronger than those previously considered by the geolocation literature. Much of the work on geophysical identification has ignored adversarial behavior: servers generate packets or other trace evidence whose measurement is presumed to reflect reality. Data sovereignty considers a more challenging scenario, as active adversaries may act strategically to fool or confuse the querier. Recent work in *position-based cryptography* considers a similarly strong type of adversary—capable of breaking nearly all previous geolocation strategies—that is able to clone itself at multiple, specific, hidden locations [5]. Capkun *et al.* present a slightly weaker adversary that is unable to locate certain landmarks ("hidden, mobile base stations") during the protocol, and thus unable to execute certain attacks [4]. Although these adversarial models were originally posed in a wireless setting, such active adversaries are closer to those we consider for data sovereignty, and make a good starting point for our future analysis.

Data sovereignty cannot guarantee that additional copies of data are not instantiated outside of a prescribed geographic area, only that there exists at least one copy of the data at an interrogation point. Considering adversaries that hold copies of the data at multiple geographic locations is outside the scope of data sovereignty. Tracking all copies of data, without total control of the network, is a different and very hard problem. We note that, from an economic perspective, it may be punitively expensive for a storage service to replicate digital archives that are large relative to existing network bandwidth, and therefore it may be infeasible to successfully answer random challenges on large data sets by selectively copying or distributing the archive to multiple geographically-distant locations. The premise of data sovereignty is that an adversary may have some incentive for re-locating its data in breach of a contract, but storing copies at multiple locations undermines any such motive.

4 The State of the Art

Tools to actively monitor real cloud performance or SLA compliance—such as *CloudCmp* [14], *SLAm* [20] or Nimsoft's commercial monitoring service—do not yet offer support for checking compliance with respect to data durability or location clauses of an SLA. Most tools do monitor certain QoS metrics potentially relevant to inferring geolocation and data presence, such as up-time and end-to-end response times. Thus, extending support to monitor data sovereignty is quite natural. A data sovereignty protocol needs to achieve two things, simultaneously: (1) proof of the physical location of a server on a network within some acceptable margin of error, and (2) proof that a client's data is indeed stored at this location. We summarize applicable technologies, and describe their relationship to our problem.

Internet geolocation

Geolocation of servers on the Internet is currently achieved through a variety of evidence-gathering prac-

tices, including mining data from *whois* databases and DNS records, using modern Internet topology tools and through the manual inspection of Internet artifacts (*e.g.* confirming a webpage is written in Chinese). These methods provide a "best guess" based on a small constellation of heuristic evidence, generously assumed to be non-malicious. The only reliable, technical method for bounding location on the Internet, however, is active measurement (*i.e.* delay probes from known landmarks) in conjunction with topological information (*e.g.* from path probing and BGP routing views) [9, 11, 13, 17]. Established commercial SLA monitoring services provide natural partners for outsourcing data audits or for acting as semi-trusted landmarks capable of participating in data sovereignty protocols.

Multiple measurements mitigate variable sources of observed delay, such as congestion, while transmission and processing delay are assumed negligible relative to propagation time. By using multiple landmarks with known positions, delay measurements allow for triangulation of the destination's feasible region. However, the correlation between delay and distance is not always strong due to Internet peering points, topology, and layer-2 traffic engineering [19]. In particular, Internet delays are known to violate the triangle inequality. This is especially true considering the power of an adversarial node against these types of measurement [8].

The network measurement problem for data sovereignty, however, differs from general IP geolocation in important ways. An adversary can only increase a landmark's observed delay, and only at the edge. This allows an adversary to "move," but the attempted move's error is constrained by the set of available landmarks. Thus, one can prove a target resides *within* some bounding area, and employ multiple landmarks to constrain the size of that area. While servers can pretend to be outside the bounding area, they may never defy the speed of light to claim falsely to be inside the bounding area (crafting delays to appear outside these borders serves no useful purpose to our adversary).

Provable data possession

Beyond the limitations of geolocating an IP address, there currently exist no techniques that effectively (let alone securely) bound the geographical location of some *data* stored in the cloud. (To our knowledge no current cloud storage providers, by themselves, provide any technical means for proving either the authenticity or the location of stored data.) A class of related technologies—which we describe collectively as *provable data possession* (PDP)—can be used to efficiently audit remote data stores, without requiring the client or the server to retrieve the entire file [3, 7, 10, 18]. PDP, however, only provides proof of the existence of data, not its location.

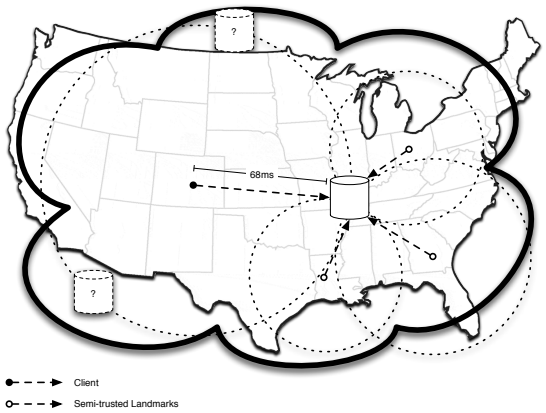

Figure 1: An initial approach to data sovereignty.

Combining the concepts of PDP with Internet geolocation to establish a novel data sovereignty protocol is non-trivial and provides a new and interesting setting for both problems. Naïvely composing latency-based geolocation with provable data possession, *i.e.* applying each technique serially and independently, provides limited assurance. Doing so establishes only two, disconnected facts: first, an unmodified copy of the data exists *somewhere* and second, the replying server exists within some known physical boundary. We attain no strong binding between the location and the data. In particular, the geolocated server may be proxying, *i.e.* relaying, the PDP challenges to some server at a different, remote location.

5 An Initial Approach

One promising strategy for building a meaningful data sovereignty protocol is to bind network geolocation query responses with some proof of data possession. Among existing PDP techniques, the MAC-based PDP scheme [10, 16] is an attractive candidate as it requires no server-side computation during the protocol; the server merely retrieves challenged blocks from storage. Our initial approach therefore considers leveraging a MAC-based PDP (MAC-PDP) as the interrogated server incurs no computational delay, thereby permitting a multilateration-style network delay measurement.

In MAC-PDP, a client breaks a file into blocks and tags each block using a message authentication code function, such as HMAC with key k. The client stores the blocks and tags in the cloud, retaining only k. To challenge possession, the client chooses c random indices, and requests the corresponding blocks and tags; the audit's probabilistic guarantee is a function of c. To verify, the client recomputes tags from the received blocks, and compares these against the retrieved tags.

To achieve data sovereignty, one could augment the MAC-PDP scheme with network delay measurement capabilities, quantifying the time it takes for the server to respond to each challenge. This basic strategy—associating data storage with a quality of service metric—has the effect of confounding two (previously orthogonal) ideas: remote file possession and service responsiveness. We note this may not be appropriate for providers whose storage guarantees come at the cost of variable, possibly long, access times; consider, for example, the seek times associated with random access using tape storage. However, imposing these additional QoS requirements on the service provider may be acceptable in many scenarios, and is reasonable to consider as an initial approach.

By combining the responses of a network delay-based measurement geolocation protocol with a PDP response, the objective is to provide a strong binding between network location and data location. Of course, one must avoid introducing any variable overhead to the server's processing of the request, so the measured latency almost entirely reflects propagation cost. MAC-PDP allows a single challenger to use the measured delay to calculate a radial distance of the responder (*e.g.* using only the client in Figure 1). A well placed challenger may be able to provably place data within the continental United States, efficiently satisfying many of the sovereignty issues addressed in Section 2.

For finer location granularity, a challenger may employ the help of friendly, semi-trusted landmarks (*e.g.* using all landmarks in Figure 1). Each landmark may challenge the server to respond to some subset of the c randomly chosen blocks, measure the response times, then authentically report the responses and measurements to the client, to aid in estimating the target's location.

As in PDP, an adversary may successfully answer challenges, while storing only part of the file locally; however, new soundness arguments are warranted. It may be possible to transfer remote blocks during an audit, so that a block that appears to be stored locally at time $t + 1$ might not have been stored locally at time t. This is comparable to giving a typical PDP adversary the ability to recover deleted file blocks, at some cost. Consider a simplistic model in which the server may transfer one remote block during the time taken to respond to a challenge, *i.e.* the number of remote blocks after t challenges is $d_t = d_0 - t$, for $t \leq d_0$. For $d_0 = 1$, an adversary may answer an initial challenge with high probability and, then, answer all future challenges with absolute certainty. In general, the soundness error in this model is bound by $\left(\frac{n-d_0+c}{n}\right)^c$, using c challenges for an n-block file. Of course, an analysis that models time and transfer costs more realistically is necessary.

The proposed approach has the advantage of requiring no computation at the server and can be immediately implemented given existing cloud infrastructure. The scheme's simplicity, however, comes with a relatively high communication cost: using block size b, at least $c \times b$ bytes must be transferred. Using more complex techniques—*e.g.* simultaneous challenges, compressing the responses using homomorphic signatures, as done by some PDP schemes [3, 18]—seems difficult: complex server-side operations add variance to the perceived delay; treating this as error adds opportunity for mischief.

6 Discussion

Data sovereignty opens a rich, new problem space with many open questions, each requiring further study. For example, what is the right way to establish and position known and trusted landmarks? Is this a role that can be played by the government? Can competing storage service providers be incentivized to act as landmarks for their competitors? Can we create a web-of-trust given some number of known (or unknown) honest landmarks?

Data sovereignty provides an explicit tool to break a level of abstraction provided by the cloud. The idea of having the abstraction of the cloud when we want it, and removing it when we don't, is a powerful one. Doing it the "right way" may require significantly different assumptions and architectures than currently exist. It would be desirable to attain data sovereignty without imposing *any* QoS requirements, perhaps leveraging some new assumptions. Does data sovereignty become easier using an overlay network of trusted, fixed-position BGP nodes to sign traffic? Can we constructively leverage un-clonable, tamperproof devices operating on-site at the storage service provider, binding *computation*, rather than data, to a location? Ideally, establishing such an assumption would isolate what makes data sovereignty across the Internet "hard," and may provide alternative strategies for tackling the problem. While some assumptions may be unrealistic to deploy universally across the Internet, it may be practical to satisfy them among the smaller population of servers and clients for whom data sovereignty is important.

When technical solutions fail to solve "hard" problems, legal remedies are often prescribed to discourage malfeasance and to punish those, *a posteriori*, who are discovered to have acted in bad faith. As such, there are legitimate questions and ambiguities about the legal protections available to data stored in a geography-agnostic cloud. For example, in a recent report [15], Microsoft raises concerns that Fourth Amendment search and seizure protections may only apply to data physically stored in the United States, belying a consumer's expected or inherent right to privacy.

Indeed, there exist many laws that govern the flow and storage of data across national borders, including legislation governing privacy law, intellectual property law, law enforcement regulations, e-discovery obligations and intelligence gathering regulations. In Nova Scotia and British Columbia, most personal data held by public bodies cannot be moved outside the borders of Canada. Australia's National Privacy Principle #9, concerning transborder data flows, prohibits the transfer of personal information to a foreign country unless certain criteria are met, including the condition that the foreign country upholds law substantially similar to the National Privacy Principles [2]. Likewise, the EU Data Protection Directive broadly restricts the flow of personal information from within Europe to any country whose domestic laws do not provide an "adequate level of protection" [1]. While the US Dept. of Commerce has organized a voluntary mechanism for US companies to certify compliance with this EU directive (the US-EU Safe Harbor Principles), the sufficiency of these mechanisms has been the subject of regular criticism [6]. In April 2010, German data protection authorities issued a resolution requiring extra diligence for German data exporters interacting with US Safe Harbor-certified entities—effectively calling into question the sufficiency of the Safe Harbor program to meet EU guidelines—holding exporters liable for lack of diligence, to face possible sanctions. Other nations have expressed reservations about data stored in US-based clouds falling under the jurisdiction of US laws like the Patriot Act.

Within the US, regulations concerning data management—including HIPAA, HITECH, GLBA, SOX, and FISMA—do not specifically regulate the physical location of stored data, although an organization's compliance and security planning may restrict location as part of its strategy. Risk management and data security analysis may be based on the properties of a particular data center: safeguards at that center, who has access, if employees hold clearances, the type of monitoring performed by on-site security personnel, *etc*. Moving data to a new location may change these analyses, leaving customers non-compliant. Additionally, organizations using remote storage for sensitive information or intellectual property may desire those data be stored at locations within U.S. borders to avoid the complications of handling data breaches or navigating legal protections (or lack thereof) in foreign nations. While many of these issues may ultimately be solved only by the courts or through legislation (matters of law), data sovereignty may be a useful legal tool in establishing meaningful evidence in future litigation (matters of fact).

References

[1] Directive 95/46/EC of the European Parliament of the Council ("Data Protection Directive"), 1995. Available at `http://bit.ly/5eLDdi`.

[2] Privacy Act 1998 (Cth) ("Privacy Act"), Schedule 3, 1998. Available at `http://bit.ly/erDc0B`.

[3] ATENIESE, G., BURNS, R., CURTMOLA, R., AMD LEA KISSNER, J. H., PETERSON, Z., AND SONG, D. Provable data possession at untrusted stores. In *Proceedings of the ACM Conference on Computer and Communications Security* (2007).

[4] CAPKUN, S., CAGALJ, M., AND SRIVASTAVA, M. Secure localization with hidden and mobile base stations. In *Proceedings of Conference on Computer Communications* (2006).

[5] CHANDRAN, N., GOYAL, V., AND OSTROVSKY, R. M. R. Position based cryptography. In *Proceedings of the International Cryptology Conference* (2009).

[6] CONNOLLY, C. US safe harbor - fact or fiction? *Privacy Laws and Business International 96* (December 2008).

[7] DESWARTE, Y., QUISQUATER, J.-J., AND SAÏDANE, A. Remote integrity checking: How to trust files stored on untrusted servers. In *Proceedings of the Conference on Integrity and Internal Control in Information Systems* (2003).

[8] GILL, P., GANJALI, Y., WONG, B., AND LIE, D. Dude, where's that IP? Circumventing measurement-based IP geolocation. In *Proceedings of the USENIX Security Symposium* (2010).

[9] GUEYE, B., ZIVIANI, A., CROVELLA, M., AND FDIDA, S. Constraint-based geolocation of Internet hosts. *Transactions on Networking 14*, 6 (December 2006).

[10] JUELS, A., AND KALISKI JR., B. S. PORs: Proofs of retrievability for large files. In *Proceedings of the ACM Conference on Computer and Communications Security* (2007).

[11] KATZ-BASSETT, E., JOHN, J. P., KRISHNAMURTHY, A., WETHERALL, D., ANDERSON, T., AND CHAWATHE, Y. Towards IP geolocation using delay and topology measurements. In *Proceedings of the Conference on Internet measurement* (2006).

[12] KUNDRA, V. Federal cloud computing strategy, February 2011.

[13] LAKI, S., MATRAY, P., HAGA, P., CSABAI, I., AND VATTAY, G. A detailed path-latency model for router geolocation. In *Proceedings of the International Conference on Testbeds and Research Infrastructures for the Development of Networks Communities and Workshops* (2009).

[14] LI, A., YANG, X., KANDULA, S., AND ZHANG, M. CloudCmp: Comparing public cloud providers. In *Proceedings of the Internet Modeling Conference* (2010).

[15] MICROSOFT CORPORATION. Building confidence in the cloud: A proposal for industry and government action to advance cloud computing. Tech. rep., Microsoft Corporation, January 2010.

[16] NAOR, M., AND ROTHBLUM, G. N. The complexity of online memory checking. *Journal of the ACM 56*, 1 (2009).

[17] PADMANABHAN, V. N., AND SUBRAMANIAN, L. An investigation of geographic mapping techniques for Internet hosts. In *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications* (2001).

[18] SHACHAM, H., AND WATERS, B. Compact proofs of retrievability. In *Proceedings of ASIACRYPT* (2008).

[19] SIWPERSAD, S., GUEYE, B., AND UHLIG, S. Assessing the geographic resolution of exhaustive tabulation for geolocating internet hosts. In *Passive and Active Network Measurement*, vol. 4979 of *Lecture Notes in Computer Science*. 2008.

[20] SOMMERS, J., BARFORD, P., DUFFIELD, N., AND RON, A. Multiobjective monitoring for SLA compliance. *Transaction on Networking 18*, 2 (2010).