Plurality of Provenance in Distributed Identity

A topic paper by Joe Andrieu (joe@joeandrieu.com) for the ID2020 Design Workshop, May 2016.

The blockchain is a robust consensus-based distributed ledger comprised of permanent, non-repudiatable public records. Such a ledger would be suitable for establishing identity if common semantics and syntax can be defined for creating and interpreting records.

Unique to such an Identity system is the inevitable plurality of assertions and authors. I propose a simple set of requirements for recording and evaluating a multiplicity of irrevocable claims correlating subjects with attributes and privileges.

First, terminology:

Subject: the focus of correlation with specific **attributes** or **privileges**. An individual, agent, or entity, for which correlation is desired, i.e., for which identity is in question.

Evaluator: the agent or entity attempting to correlate the s**ubject** with a given set of **attributes** or **privileges**.

Author: the originator of a given assertion.

Assertion: A statement by an **author** correlating a **subject** with a specific **attribute** or **privilege**.

Claim: An **assertion** cryptographically signed by its **author**.

Attribute: a fact describing the **subject**, e.g., a name, or date of birth, citizenship, age, etc.

Privilege: an affordance given to a **subject** by an **author**, whether for reading, writing, executing or some other interaction.

Where **attributes** describe the **subject**, **privileges** describe services, capabilities, or allowances that should be afforded to the **subject**.

I use these terms instead of more familiar user-centric ones because although the term “relying party” is typically suitable for what I mean by the **evaluator**, “Identity Provider” conflates the service that might serve **claims** with the **author** that creates them. In particular, when using the blockchain as an identity ledger, it can be confusing which component in the solution correlates to traditional “Identity Providers:” the ledger, a wallet, the miners, or something else.

Now to the topic…

Single source assertion of identity lacks robustness.

Anil John put it this way in his proposed "Canvas Theory of Access Control":

> *Over time, as a credential is used in transactions, the image of the credential holder becomes more and more clear on the canvas. And based on this visibility, combined with many other factors, the level of access can increase.*[1]

Jeff Jonas said it this way, in his discussion of the value of space-time travel data revealed/created by mobile phones:

> *It doesn't matter who you say you are! Where you are (space), when you're there (time), and your movements over time (travel) are closer to the truth.*[2]

In the reverse, we saw the deanonymization of users after the poorly conceived AOL Research project released keyword data for over 650,000 users in 2006.[3]

Not only can identity be discovered from the evaluation of a plurality of cumulative data, perhaps such cumulative data is *the most fundamental foundation* of identity.

It follows then, that identity can be addressed as a problem of reputation. Accumulate enough information of sufficient reliability and identity *emerges* from correlating a subject with their history:

> *We therefore define reputation as "information used to make a value judgment about an object or a person."*[4]

The blockchain provides fertile ground for establishing the foundation for such a non-repudiatable record, upon which we can build a distributed system of identity as reputation. Or as the Violent Femmes put it:

> *This will go down on your permanent record.*[5]

Such a distributed ledger would allow any evaluator to use their own algorithms, judgment, and supplemental (proprietary) information for resolving identity, while allowing any author the opportunity to record their own assertions as a matter of public record. It would be the world's first truly distributed, customizable identity fabric.

---

[1] John, Anil, "Canvas Theory of Identity LOA vs Canvas Theory of Access Control", 2011-06-12. Online. Accessed 2016-05-16

[2] Jonas, Jeff. "Your Movements Speak for Themselves: Space-Time Travel Data is Analytic Super-Food" 2009-08-16. Online. Accessed 2016-05-16. http://jeffjonas.typepad.com/jeff_jonas/2009/08/your-movements-speak-for-themselves-spacetime-travel-data-is-analytic-superfood.html

[3] Wikipedia. "AOL search data leak" Online. Accessed 2016.05.17 https://en.wikipedia.org/wiki/AOL_search_data_leak

[4] Farmer, Randy and Glace, Bryce, "Web Reputation Systems and the Real World", *Building Web Reputation Systems: The Blog*, 2014-06-16. Online. Accessed 2016-05-16. http://buildingreputation.com/

[5] Gano, Gordon, "Kiss Off" *Violent Femmes.* Vinyl, LP, Album. Gorno Music. 1983.

To realize this possibility, four capabilities should be established.

1. Data integrity
2. Data authority
3. Assertion authority semantics
4. Deep provenance

1. **Data integrity**. Can we verify that a given assertion was actually made by its alleged author? We accept that cryptographically signed claims provide sufficient confidence that the signed assertion is, generally, an assertion of the author who signed it. In the relatively rare case of compromised keys, we must rely on deep provenance for evaluators to respond dynamically to knowledge of such compromises. Signed claims recorded on the blockchain provide further evidence that a particular assertion was irrevocably made at a given moment in time. Even if that key is later compromised, those claims recorded while it was still valid retain their integrity. The older a claim on the chain, the greater the probability it *must* be authentic.

2. **Authorial merit**. Is the author recognized as a valid authority for the assertion being made? This depends the judgment of the evaluator for the assertion in question. In some cases, the root authority for a class of assertions can make a definitive claim. For example, the U.S. Department of State can definitively claim that a particular set of identifiers and attributes correlate to a U.S. Citizen, or a credit card company can definitively state the credit limit at a specific point in time for a specific card number. This is the most fundamental authority. Other times, the author may be accepted as an authority because the evaluator's relationship with the author is sufficiently strong relative to the assertion. For example, we might accept that "John Smith" at a johnsmith@example.com is the "John Smith" we went to high school with because a mutual friend gave us that email address with that assurance. Fundamentally, the evaluator must decide which authors it trusts for different assertions. The records must provide enough structured information to evaluate the nature of an assertion and, potentially, to use the ledger history to monitor the veracity and reputations of specific authors.

3. **Assertion authority semantics.** To judge a given claim, evaluators need to know the basis upon which the author made the assertion. How does the author know what they are asserting is true? An assertion by a root authority is fundamentally different than an assertion by an identity proofing bureau which has independently verified that the subject in question presented physical documentation, in person, to a sworn agent. Both of which are different from an assertion based on information submitted on a web form. All of these statements should be possible using semantics that make it clear to evaluators how to treat the assertion in their correlation algorithm.

4. **Deep provenance.** A public ledger where any author may present claims about any subject will, by its nature, eventually present a multiplicity of assertions for any given subject. To resolve potential conflicts or ambiguities in assertions from a variety of sources—which themselves may vary in quality over time—an identity system should retain the provenance for every assertion in the system.

Every assertion comes from somewhere. Today, most systems retain extremely shallow history of attributes. For example, OAuth identity providers present attributes like email and name, with no clear explanation of how any why the provider believes those to be valid; in turn, many relying parties accept those attributes as "fact" without keeping track of the source of the original assertion. In order to dynamically respond to compromised keys, evolving knowledge of bad actors, and the accretive trust

that can accrue from a long term history of consistent behavior, even assertions as simple as the name or birth date of an individual should retain the sources from which those assertions have been received. Such provenance also allows for social notions of identity to emerge, where large numbers of moderately trusted authors—or the triangulation of just a few highly trusted social relationships—can be used to dynamically assure identity for certain uses. The ledger provides permanent retention of claims; evaluator software should as well.

Requirements for a Plurality of Provenance Identity System

I suggest that a distributed identity ledger support signed claims with the following meta-structure, whether recorded on the chain or recorded by cryptographic reference and stored in a distributed hash table (DHT):

1. One or more assertions correlating a subject with an attribute or privilege
2. Statement of the claimed authority of the author for that assertion, e.g.,
    a. Root authority
    b. In-situ physical observation (Author physically observed the subject)
    c. In-person verification of third party documentation
    d. Self-reflective linking, e.g., symmetric rel-me tags
    e. Self-asserted via REST API or form submission
3. Optional context supporting the assertion, e.g.,
    a. Date of the assertion
    b. For a web form: the browser, IP address, browser fingerprint, server name, server version, SSL identifiers, etc.
    c. For a verification assertion: the documents presented and the agent who reviewed the documentation

And that any system for evaluation

4. Retain all provenance information
5. Have an algorithm for aggregating and correlating multiple, diverse claims potentially related to a candidate subject
6. Be capable of dynamically responding to changes in the trustworthiness of sources, additional records, and other forms of refutation or concurrence

The specific lexicon for assertions, authority, and contexts should be both standardized and extensible, based on requirements. The specific representational model (abstract schema, e.g., RDF) and data format (e.g. encoding as JSON or RDF-XML) are engineering projects to be resolved. Beyond a baseline reference implementation, algorithms for resolving identity can be developed and customized by vendors and application developers to meet real-world needs of service providers.

As long as the records in the ledger utilize a consistent and stable representation, a multiplicity of author and evaluator tools can be developed for a broad range of applications. As long as that representation is semantically extensible within a consistent syntax and encoding, the type of data recorded can evolve to meet future needs.