

*SEMINAR REPORT ON*

**Secure Scuttlebutt: An Identity-Centric Protocol**

**for Subjective and Decentralized Applications**

*SUBMITTED TO*

**Department of Computer Science & Engineering**

*by*

Chinmay Sanjay Deshpande

Registration Number 170905432

Roll Number 50

Semester VII Section D

Name & Signature of Guide

Name & Signature of Evaluator 1 Name & Signature of Evaluator 2

(August 2020)

|  |  |  |
| --- | --- | --- |
| **Table of Contents** | | |
|  | | |
| **Section 1** | **INTRODUCTION** | **1** |
| **1.1** | General | 1 |
| **1.2** | Motivation | 1 |
| **1.3** | Structure of the Report | 2 |
|  | | |
| **Section 2** | **BACKGROUND THEORY** | **2** |
| **2.1** | Information-Centric Networking Approach | 2 |
| **2.2** | Gossip Protocol | 4 |
| **2.3** | Selective complete log replication and Subjective Reading | 4 |
| **2.4** | SSB Architecture and Protocol | 5 |
|  | | |
| **Section 3** | **METHODOLOGY** | **7** |
| **3.1** | Predicting human traits and attributes by analyzing digital behavior | 8 |
| **3.2** | Running Distributed Applications over SSB | 9 |
|  | | |
| **Section 4** | **RESULTS AND DISCUSSION** | **10** |
| **4.1** | Observations from Human Attributes Prediction Study | 11 |
| **4.2** | Benefits of SSB | 13 |
|  | | |
| **Section 5** | **CONCLUSIONS AND FUTURE ENHANCEMENTS** | **14** |
| **5.1** | A Decentralized Approach | 14 |
| **5.2** | SSB as an alternative and future challenges | 14 |
|  | |  |
| **REFERENCES** | | **15** |

1. **INTRODUCTION**
   1. **General**

Secure Scuttlebutt (SSB), a peer-to-peer, event sharing-protocol and architecture for social apps is a project which started in 2014 during the “decentralized web movement”. it underwent several protocol design and implementation cycles, and it currently stands as a stable service providing over 10,000 users with a variety of media communication applications along with end-to-end encryption(cryptographic protection) and data integrity(privacy of metadata) in pure peer-to-peer network mode. It follows the Identity-Centric Networking Architecture, as opposed to the Information-Centric approach used in major social networking application networks which have a centralized-governing data-driven design rationale.

* 1. **Motivation**

Facebook is the biggest social media site, with more than 2.5 billion Monthly Active Users (MAU). There are more than 65 million businesses using Facebook Pages and more than six million advertisers actively promoting their business on Facebook. The Facebook–Cambridge Analytica data breach which unfolded in March 2018, shed light on the data leak incident whereby millions of Facebook users' personal data was harvested without consent by Cambridge Analytica, predominantly to be used for political advertising.

Big tech giants such as Facebook, not only monopolise the market and defeat the purpose of the principle of Net Neutrality, but could also be used to collect personal information and even influence decisions by displaying content specifically tailored to target a group of people by analysing their data. This, combined with fake news advertising allegations, question the data security and integrity of centrally organized social applications present. There is little, or no transparency between the users and the methods in which their data is collected and analysed – the users become consumers of the information they are subjected to, based on systematic storage and extensive study of their own personal data without knowledge.

SSB may prove to be a suitable alternative approach while designing new social media applications, which will only focus on the main objective – sharing data among selected peers with confidentiality and security. The apps can then form a virtual environment where no organization/body centrally governs and manipulates the personal information passed over the network. The main features of this protocol include the secure, append-only log data structure, and the selective log replication based on the ‘social-graph’ created, which dictates the data flow among the peers in the network.

* 1. **Structure of the Report**

**Section 2** briefly describes the currently used Information-Centric Networking approach. It further gives an overview of the basic underlying technologies, such as the Gossip Protocol, Selective complete log replication and subjective reading (how they are implemented). Finally, the SSB Protocol and Architecture is covered in detail.

**Section 3** discusses the study for finding human traits and attributes based on the digital records of their behaviour, focussing on Facebook likes. It also demonstrates how applications can be built over the SSB architecture.

**Section 4** extrapolates the results of the previous study and forms a structured argument in favour of decentralized approach and architecture designs, such as SSB.

**Section 5** presents the conclusion of the report, urging the necessity of transforming the current Centralized Network design to a Decentralized one. The research opportunities in SSB protocol and architecture design follow.

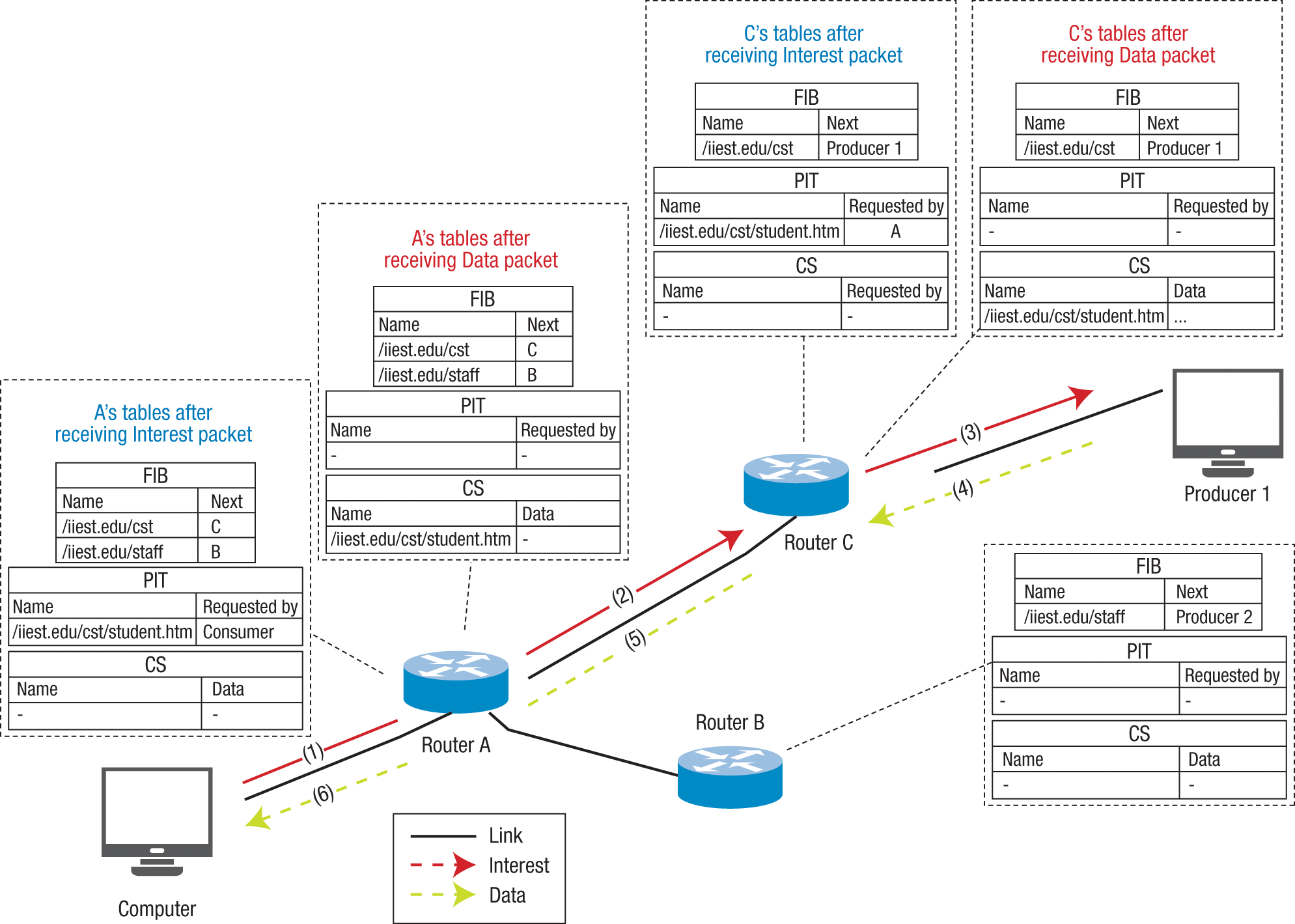
1. **BACKGROUND THEORY**
   1. **Information-Centric Networking Approach**

Information-Centric Networking (ICN) is based on a receiver-driven content retrieval model, or a *pull-model*. The user will retrieve information based on the requested content and not based on a specific location where the content is stored. It involves developing a naming scheme, which would identify all the content uniquely, provide security (with the help of self-certification or attestation by third party), be scalable, be easily translated by the content providing device and supports caching.

ICN considers 4 important characteristics for name-based routing - content-oriented, robustness, efficiency, and scalability. Regarding caching, it uses local information to determine what to cache. Any node in the network can act as an edge-node cache at any time, which means that the existing network can be made into content distribution networks since the content requests do not rely on location or host. There are many implementations of ICN.

One such project, built using the principles of CCN (a data named communications architecture with packets being addressed with content names; built in Palo Alto research centre, 2009), is Named Data Networking (NDN). It uses hierarchical naming scheme, where a client requests content by sending an Interest packet with the content's name, which is then sent to one of the Content Routers (CR) present in the network, each having a Forwarding Information Base (FIB), Pending Interest Table (PIT), and Content Store (CS). The FIB stores pairs of names to forwarding output direction. This is so that Interest packets can be forwarded to the correct place based on the requested content name. The PIT stores pairs of names to which interface requested the content. This is then used to backpropagate data objects back to the subscriber. Lastly, the CS is storage for content that travels through the interface. This is a local cache. Once an Interest packet arrives at the CR, the FIB, PIT, and CS in the CR are updated.

One of the objectives of NDN is to optimize data distribution for large content providers while assuring data/content integrity. NDN currently achieves those aims by: (1) initiating data transfers *after* the interested users are known by the network (*pull-model*), (2) utilizing existing certificate infrastructure for authentication, and (3) deploying a naming scheme that reflects the hierarchical organization of major content providers, such as universities, governments, and major media companies. The NDN architecture can be observed in Figure 1.



**Figure 1: NDN Architecture**

NDN overcomes many shortcomings of IP, such as the problem of address space exhaustion, address management, NAT traversal, multicasting in the network, and high loss end-to-end reliability.

* 1. **Gossip Protocol**

Anti-entropy, or gossip, is an attractive way of replicating state that does not have strong consistency requirements. With few limitations, updates spread in expected time that grows logarithmic in the number of participating hosts, even in the face of host failures and message loss. Gossip protocols purport to deliver messages within a certain configurable number of rounds with high probability, and thus provide synchronous guarantees. It has capacity that is limited by available bandwidth for transporting gossip data and CPU cycles for generating and processing the gossip messages.

There are two classes of gossip: anti-entropy and rumour-mongering protocols. Anti-entropy protocols gossip information until it is made obsolete by newer information and are useful for reliably sharing information among a group of participants. Scuttlebutt utilizes the anti-entropy protocols for selective log replication, through the process of reconciliation.

SSB organizes information to be delivered around *named data streams* for signed events, where the basic addressing unit is a fulllog which may produce new event messages in the future. Thestreams guarantee reliable causal ordering and authenticity. Delivery of these streams follows the *push-model*: once a peer wishes to access a stream of their interest, new updates are transferred automatically without being requested individually. Flow-control (back-pressure) in the current overlay implementation is done implicitly by the TCP connections used to deliver data among peers.

* 1. **Selective Complete Log Replication and Subjective Reading**

SSB relies on the core insight that each peer is only interested in a subset of the global data pool, thus it is feasible to locally store all the data a participant is interested in. It does so by partitioning the data pool such that all data is associated with the identity that produced it. Participants select specific data from this pool by identifying the set of identities the data belongs to. This creates a *“social graph”* along whose edges data flows (Figure 2). In the event of an overall system scaleup, the amount of data any single peer is interested in and thus needs to be handled by the network does not change much.

A picture containing table, drawing

Description automatically generated

**Figure 2: SSB’s “Internet of Identities” – Users A, B and C replicate logs (a, b, c, . . .) based on whom they follow: C does not follow A, hence has no log a. A and B follow each other such that when A follows C, A will get C’s log c via B - new content is pushed directly if possible and through intermediary friends if necessary.**

Each participant can publish data to their single writer, append only log. This choice of data structure allows efficient replication and verification of the integrity of received data.

Because replication in SSB is guided by a peer’s social graph and, as a result, is selective, different end devices will have access to different sets of log replicas, leading to different views of the world, which is referred to as the *“subjective reader*” approach. This property of SSB is considered desirable: each peer is free to consider data sources of its own choosing instead of having to feed from a centrally provisioned or otherwise converged view, i.e. the peers will only see the log records of their social circle instead of the entire data flux from the global data pool.

While it is possible to implement consensus protocols over SSB, or to designate central data aggregators from which many peers consume the consolidated outputs, the SSB network itself deliberately does not offer consensus services nor central content (directories etc), since the main objective is to only receive the requested content.

* 1. **SSB Architecture and Protocol**

Each user in SSB has a keypair to identify the *author* of the logs by the means of the private key. This keypair can be randomly generated by the users, thereby removing the need of a central authority for the introduction of new users in the network.

All the nodes running the SSB Protocol act as *relays,* which have the main objective to connect with other relays and exchange log updates, along with point-to-point encryption and updated message validation. Beyond that, it offers an API to its peers. Peers can host arbitrary programs that issue remote procedure calls (RPCs) to the relay. This functionality includes appending to a log (if you are the author, i.e. if you know its private key), reading from logs, requesting which logs a relay should replicate, and fetching BLOBS and out-of-order messages.

The *single-writer append-only logs* of SSB consist of entries (called messages) that include a backlink in the form of a cryptographic hash of the previous message (or a special indicator for the first message of a log). The most distinguishing feature of this linked list, when compared to a regular blockchain, is that each SSB user maintains their own log and cryptographically signs all their (and only their) messages.

Concretely, each message (Figure 3, contains the following information about data:

• The *backlink* to the previous message, or a null value

• The *public key* of the message’s author

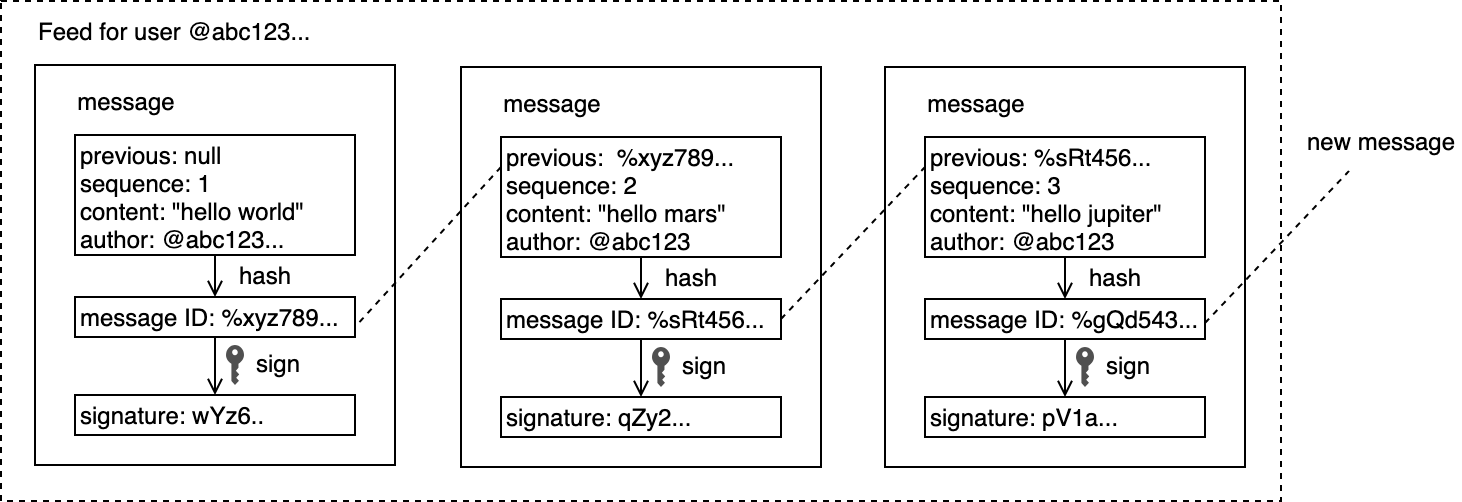
• The *sequence number* of the message (which must be one more than the sequence number of the previous message, or exactly one if it is the first message of the log)

• A *claimed timestamp* of when the message was created

• A *hash indicator* that specifies the concrete hash function that was used to compute the backlink

• The *content* of the message

• The *author’s signature* over all the previous data



**Figure 3: SSB Log and Message structure**

SSB spans over 3 independent layers of protocols. The most fundamental protocol is the *message format*: all peers need to agree on what constitutes identities, valid messages, and how to compute hashes to address messages and BLOBS. This constitutes the “thin waist” of SSB (Figure 4).

Next is the specific mechanism by which relays exchange data. The default Remote Procedure Call (RPC) mechanism is one option, but alternative mechanisms such as distribution via a sneakernet could also be used. Different users/nodes who do not follow a common replication mechanism could still interact indirectly in the presence of some relays that understand multiple replication protocols.



**Figure 4: SSB Protocol Stack**

Therefore, the core logical replication protocol by which a relay serves its clients is fully independent from the actual dissemination protocols. Lastly, the publishing and interpretation of application data in such messages is taken care of by the layer on top of the thin waist.

1. **METHODOLOGY**

This section discusses the quantitative and qualitative techniques used to predict human traits and personal attributes by analyzing digital behavior records, focusing on Facebook likes. We further use a demonstrative approach by running a distributed application over SSB and inspect the compatibility of the protocol, the data structure used, and the replication procedure used in SSB with such applications.

* 1. **Predicting human traits and attributes by analysing digital behaviour**

This study was conducted using easily accessible digital records of behaviour, Facebook Likes, to automatically and accurately predict a range of highly sensitive personal attributes including sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender.

The analysis is based on a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric tests (Figure 5). The proposed model uses dimensionality reduction for pre-processing the Likes data, which are then entered into logistic/linear regression to predict individual psych demographic profiles from Likes.

It has been shown that age, gender, occupation, education level, and even personality can be predicted from people’s Web site browsing logs. Similarly, it has been shown that personality can be predicted based on the contents of personal Web sites, music collections, properties of Facebook or Twitter profiles such as the number of friends or the density of friendship networks, or language used by their users.

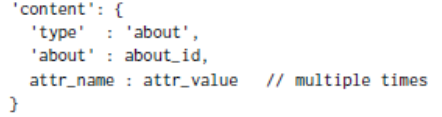


**Figure 5: The design of the Study**

The f selected traits and attributes that reveal how accurate and potentially intrusive such a predictive analysis can be, including “sexual orientation,” “ethnic origin,” “political views,” “religion,” “personality,” “intelligence,” “satisfaction with life” (SWL), substance use (“alcohol,” “drugs,” “cigarettes”), and basic demographic attributes such as “age,” “gender,” “relationship status,” and “size and density of the friendship network.” *Five Factor Model* personality scores were established using the *International Personality Item Pool (IPIP) questionnaire* with 20 items. Intelligence was measured using *Raven’s Standard Progressive Matrices (SPM)*, and SWL was measured using the SWL Scale. Age, gender, relationship status (“single/in relationship”), political views (“Liberal/Conservative”), religion, and the Facebook social network information were obtained from users’ Facebook profiles. Users’ substance abuse was recorded using online surveys. Ethnicity was determined based on visual inspection of profile pictures.

* 1. **Running Distributed Applications over SSB**

The replication model SSB incorporates makes many distributed/collaborative applications to be implemented easily by abstracting most of the complexity regarding the distribution of the updates. There are multiple applications written as well as modified by developers and contributors to be implemented over SSB following the protocols and replication. One such application can be SSB user directory, which is the user database associating cryptographic IDs with human readable attributes.



**Figure 6: (a) User Directory data structure; (b) Message format**

The message format (Figure 6(b)) is used to identify all messages of type ‘about’, and the database is updated along with the message’s most recent attributes value using the directory data structure (Figure 6(a)). ‘about\_id’ is accompanied with key/value pairs prepended by publishing authors ID, ‘author\_id’.

Analysing with the CRUD actions – Creation happens once a new SSB identity updates their own *about* message to their log; The above data structure is used for Reading the database; A new *about* message is added to the logs of the author and the peers databases during Updating; Deletion is not possible unless the user blocks a particular user ID as well as the IDs which have written an update for it.

Another implementation of SSB is the use of replicated logs to support ‘infrastructure-less’ distributed systems, which only work due to each peer being able to locally provide the updates to user requesting it; the logs statuses stored by each peer are indexed, aggressively cached, and a map-reduce strategy is used to find the latest application status.

One problem which can be observed is the need of full re-evaluation of the reduce function every time the index is updated/changed. Taking a scenario where a new identity is added to the set of followed identities of a user A, all its log posts and new updates need to be available, hence integrated into the database of A in correct order.



**Figure 7: Example of extracting application data spread over multiple collaborating logs and dealing with not-yet delivered data.**

A simple adopted solution is to use timestamp claimed by the author’s post, so that the time-sorted list can then be modified by adding new messages in a chronological order by the reduce function performing a topological sort based on the causality relationship with other posts and their replies (Figure 7).

1. **RESULT AND DISCUSSION**

This section puts forth the result of the Human traits prediction study and the performance of SSB based distributed applications. The findings of the study include statistical representation of the prediction model and its success. The analysis of SSB applications, the problems they have resolved, and their shortcomings follow next. We then discuss some benefits of SSB over the traditional networking architecture and mention areas which may need further research.

* 1. **Observations from Human Attributes Prediction Study**

The following graphs display the prediction accuracy of the various types of attributes after the subjection of Facebook Likes and other surveys (user data) to a Linear Regression Model. The traits and personal attributes are divided into dichotomic and numeric types for the result (Figure 8 and 9 respectively); Figure 10 shows the accuracy of traits based on the number of Likes available for analysis.



**Figure 8: Prediction accuracy of classification for dichotomized attributes**



**Figure 9: Prediction accuracy of regression for numeric attributes and traits**

These results reveal that a wide variety of people’s personal attributes can be automatically and accurately inferred using their Facebook Likes. Similarity between Facebook Likes and other widespread kinds of digital records, such as browsing histories, search queries, or purchase histories suggests that the potential to reveal users’ attributes is unlikely to be limited to Likes. On the other hand, the predictability of individual attributes from digital records of behaviour, even if incorrect, may have considerable negative implications, because it can easily be applied to large numbers of people without obtaining their individual consent and without them noticing, for the personal benefit of an organisation/conglomerate/politicians and officials.



**Figure 10: Accuracy of selected predictions as a function of the number of available Likes**

* 1. **Benefits of SSB**

By investigating the nature of Distributed Applications running over SSB, we find that replication of individual append-only logs, the core of SSB’s protocol solves numerous hard problems in Distributed Systems – 1. It requires no run-time checks/configuration management since it is a radically decentralized approach needing only minimal specification-level coordination among peers; 2. It deploys a cryptographic ID system and maintains a social graph informing routing which results in a receiver-driven approach where data only flows where it is needed and provides flexibility in the actual data dissemination strategy; 3. Every peer is a publisher by design. This property goes beyond the decentralized approaches which assume that there exist replication servers but keep the separation between a data transport network and a server layer; 4. Finally, log replication leads to a distributed system with inherent high resilience as any communicating element (message) carries a persistent copy of the data, as opposed to the traditional distributed systems, where coordinating data persistence often turns out to be an add-on task, or requires at least a special recovery service.

1. **CONCLUSION AND FUTURE ENHANCEMENTS**
   1. **A Decentralized Approach**

The aim of this report, as stated in the introduction, was to put forth the demerits and potential threats (either in the form of flaws, or malicious intent of individuals/organizations) that maybe found in the currently used Centrally-Organized data governing social networking model.

While there may still be much extensive research required in the exact implementation, standards and protocol designing, the benefits that decentralised network architecture has to offer is very appealing and it has the potential to take over the future of the internet. The internet was meant to be open and decentralized, and it is in the hope of developers that one day, this net neutral, unbiased and transparent system will be deployable and accepted globally. The peer-to-peer network mode could be an answer to government censorship and monopoly posed by the giant tech companies. As the masses garner for democracy over the internet, decentralization may be the one of the, if not the only way to achieve this. A network on its own, giving people the freedom, they desire online, could one day become a reality.

* 1. **SSB as an Alternative and Future Challenges**

The core novelty of SSB is that data replication occurs at the granularity of complete, self-certifying append-only logs of messages by an author. This approach leads to a simple, yet efficient information-centric service abstraction that lends itself well to a large class of applications. By embracing push based eventual delivery and subjective interpretation of data, SSB gets to sidestep common sources of complexity. A community of over multiple thousand users interacting through a variety of applications confirm the viability of the approach.

Focussing on identities leads to certain challenges, such as pseudonymity (user privacy), non-refutability and immutability of the user logs, handling ‘onboarding’ (introduction and accommodation of new identity among established peers and their pre-existing social graphs), non-interoperable message types (due to absence of a central coordination, which could result into a multitude of mutually non-understanding fractions using different messages or interpretations). But it enables elegant, decentralized solutions to common problems with Information-Centric systems, and hence should be considered an alternative, as well as a field with extensive research opportunity.

**REFERENCES**

1. Zhanglong Peng, (2019), “*Decentralized Internet*”, https://www.cse.wustl.edu/~jain/cse570-19/ftp/decentrl/index.html
2. Dominic Tarr, Erick Lavoie, Aljoscha Meyer and Christian Tschudin, (2019), “*Secure Scuttlebutt: An Identity-Centric Protocol for Subjective and Decentralized Applications*”, Macau: ACM Conference on Information-Centric Networking 2019.
3. B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher and B. Ohlman, (2012), "*A survey of information-centric networking*," IEEE Communications Magazine, vol. 50, no. 7, pp. 26-36.
4. Kosinski, M., Stillwell, D., and Graepel, T., (2013), “*Private traits and attributes are predictable from digital records of human behaviour”*, Proceedings of the National Academy of Sciences, 110, 5802 - 5805.
5. Van Renesse, Robbert and Dumitriu, Dan and Gough, Valient and Thomas, Chris, (2008), “*Efficient reconciliation and flow control for anti-entropy protocols*”, ACM International Conference Proceeding Series.
6. Facebook–Cambridge Analytica data scandal, https://en.wikipedia.org/wiki/Facebook-Cambridge\_Analytica\_data\_scandal