Towards End-to-End Security in Branchless Banking

Saurabh Panjwani†
Microsoft Research India
saurabh.panjwani@microsoft.com

Abstract
Mobile-based branchless banking has become one of the key mechanisms for extending financial services to low-income populations in the world’s developing regions. One shortcoming of today’s branchless banking systems is that they rely largely on network-layer services for securing transactions and do not implement any application-layer security. Recent results show that several of these systems are, in fact, not end-to-end secure.

In this paper, we make the case for designing mobile-based branchless banking systems which build security into the application layer and guarantee end-to-end security to system users. We present a threat model which captures the goals of authenticated transactions in these systems and then provide recommendations for solution design based on our model’s requirements.

1 Introduction
While western societies continue to advance and experience the benefits of modern financial services, there remain 2.5 billion people in the world without access to a basic savings account [7]. Recent efforts, under the hood of branchless banking, are beginning to bridge this gap by enabling banks and other financial institutions to operate in remote areas without incurring the cost of setting up brick-and-mortar offices or ATM terminals. Systems designed under this paradigm operate through a network of human agents who facilitate cash-based transactions for beneficiaries and use cheap and prevalent technologies for communicating with the service provider. There are tens of branchless banking systems already in existence today and in sum, they serve more than 50 million people and transact over $100 million on a daily basis [11, 10].

The deep penetration of mobile phones in developing regions makes cellular networks the preferred communication medium in branchless banking. Transactions typically involve a customer and an agent meeting in one location, sending and receiving messages to/from a bank server through their mobile phones and transferring some physical cash from one person to the other. As in any financial service, ensuring security of transactions is paramount and indeed, most of the current branchless banking systems take steps in this direction.

One shortcoming of most of today’s branchless banking systems is that they provide security at the network layer only and do not implement any application-layer cryptography. For example, M-Pesa [5], which is the pioneer of the branchless banking concept and serves over 50% of Kenya’s adult population, uses a custom-made SIM Tool Kit (STK) program to protect transaction messages exchanged between client phones and the server. Not much is publicly known about M-Pesa’s security algorithm but recent attacks on the system reveal that it does not guarantee end-to-end security to customers [16]. Other key players like G-Cash [3] in the Philippines rely directly on GSM’s default security services to protect client information but these services are known to offer very weak security guarantees [9, 18]; in [13], it is argued that vulnerabilities in GSM’s security suite could be used to subvert G-Cash transactions. Indeed, the question of what application-level security means in the context of branchless banking does not seem to be well-understood yet, neither in the academic literature nor in practice.

In this paper, we take the first step in addressing the above question. We present a protocol and threat model which captures the goals of end-to-end security in branchless banking systems. We focus on the task of authenticating transactions in this paper; ensuring transaction privacy in branchless banking, as will become evident, is fairly non-trivial and will be explored in future work. Our threat model, besides incorporating the obvious external threats on protocols (e.g., man-in-the-middle tampering), also addresses the possibility of insider attacks—attacks in which an agent or a customer behaves maliciously against other parties. For example, we assume that agents, who are proximally present during transactions, can eavesdrop on information provided by customers and can use this information adversarially.

We then proceed towards designing a system that meets the security requirement of our model. There are two
unique challenges to this design task, which we highlight here. The first is that of insider attacks: how do customers authenticate themselves to the bank given that their credentials could be easily leaked to nearby agents? Knowledge-based credentials like passwords are a weak candidate tool for this because they operate well only in scenarios where eavesdropping is not feasible. The use of biometrics, though an interesting possibility, is fraught with issues of privacy and irrevocability, which are exacerbated in the current context by the eavesdropping power available to bank agents. We thus propose the use of one-time password (OTP) based hardware tokens in our system since they seem to provide the best defense against eavesdropping and are also commonly used in practice [4]. Though OTP-based schemes have some limitations (e.g., OTP tokens are susceptible to theft), we discuss how these limitations could be addressed by carefully augmenting OTP’s with the use of fixed numeric passwords.

The other issue which makes our task difficult is the context in which branchless banking operates. These systems cater to some of the world’s poorest people, who have limited educational backgrounds and who rarely possess programmable phones. (Over 100 million people in India still seem to own a dumb phone [17].) The implications for design are: (a) use a simple interface (minimize protocol steps and human computation requirements); and (b) to the extent possible, avoid programming client phones. While we are able to reasonably satisfy the first constraint, we require at least some phones in our system to be programmed for without this, end-to-end security seems impossible to achieve. We first present our ideas to build systems assuming all phones are programmable. Later, we present variations for settings where only agent phones are assumed programmable. These variations make additional setup assumptions (e.g., greater use of OTPs) and achieve slightly-relaxed forms of security. We note that a few existing systems already assume programmability of agent phones [1], an indicator that the assumption may not be a deterrent to deployment.

2 Related Work

Despite the rife deployment of branchless banking systems, research on security issues in these systems is extremely sparse. Most of the currently-deployed systems seem to either rely on GSM’s inbuilt mechanisms for secure transactions while some develop their own network-layer schemes to do the same. The pitfalls of network-layer security design are well-documented in the literature [15] and relying on the inbuilt security mechanisms of 2G GSM, the most commonly used communication standard in branchless banking, is especially risky given the history of security exploits associated with it [9, 18].

An important case in point is M-Pesa [5], perhaps the biggest and most cited branchless banking system in the world today. In this system, each client authenticates himself to the bank during transactions using a 4-digit PIN uniquely assigned to him. (The PIN is sent along with other transaction details.) The SIM card of the client is pre-programmed with cryptographic software which purportedly protects the client’s PIN during transit. The algorithm implemented by this software is not publicly known. While such practice of obscuring algorithmic details of cryptographic software is in itself questionable, we point out two other limitations of the M-Pesa solution. First, the solution exists entirely at the network layer, which implies limited protection, if any, from network operators (e.g., network operators could have complete access to client PINs). Second, protecting client PINs alone does not imply that the system is end-to-end secure. Indeed, a recent attack on the system [16] confirms that the solution does not provide end-to-end security to clients, neither to customers nor to agents. In this attack, a malicious customer, with the help of a remote conspirator spoofing SMS’s on behalf of the bank, was able to convince an unsuspecting agent to yield cash even without the bank having recorded a cash transfer. The attack exploited the simple fact that the system does not enable clients to authenticate bank-originating SMS’s, thus making them susceptible to easy spoofing attacks [6].

3 System Model and Threats

We present a model for branchless banking systems and the threats they could be subjected to in practice. Our system model is an abstraction of well-known branchless banking implementations like M-Pesa (in Kenya) [5], GCash (in the Philippines) [3] and Eko (in India) [2].

3.1 Basics

A branchless banking system involves three types of entities: bank, customer and agent. There is a single bank \( B \) in the system and all customers and agents are its clients. The bank manages financial information for all clients. Agents are special clients who assist customers in conducting financial transactions with the bank.

Clients engage in different protocols and each protocol involves communication with the bank. Every client \( X \) owns a personal device \( D_X \) (e.g., a mobile phone) that can communicate with a server maintained at the bank.

3.2 Protocol Types

To join a branchless banking system, each client \( X \) must execute an enrolment protocol wherein he submits identification information to the bank (e.g., ration card, proof of residence) and the bank, after carrying out suitable background checks, assigns him a unique account number \( ID_X \); this could just be a unique identifier of the de-
The bank sends a request to the bank server using a protocol but this time to reduce his account balance. If the bank server validates the withdrawal message in order to authorize the bank, through a database maintained by the bank server, drop messages selectively and, possibly, inject spoofed messages (i.e., messages with arbitrary sender addresses) into the channel. Giving the adversary the ability to disrupt communication between system entities is also a practicable threat: spoofing the originating address of an SMS is known to be easy [6].

3.3 External Threats

We consider two types of attacks on a branchless banking system—external and internal. External attacks are those which disrupt communication between system entities. We assume the existence of an adversary who can eavesdrop on all communication between client devices and the bank server, drop messages selectively and, possibly, inject spoofed messages (i.e., messages with arbitrary sender addresses) into the channel. Giving the adversary the ability to disrupt communication between system entities is also a practicable threat: spoofing the originating address of an SMS is known to be easy [6].

3.4 Internal Threats

It is plausible that internal parties in a branchless banking system are sufficiently incentivized to play malice against each other. In particular, since agents facilitate transactions for multiple customers (more than a hundred, at times) and have a relatively limited amount of control exercised over them (due to their physical distance from the bank), they could easily be inclined to deviate from their prescribed behaviour. Similarly, customers, being remote from the bank, could try to violate norms and cheat either on other customers or on agents.

We thus allow certain adversarial powers to all parties...
(except the bank) in a branchless banking system. We assume that agents have two capabilities: first, they can eavesdrop on any information that is provided by customers during transactions. This captures the possibility of physically observing such information (e.g., shoulder-surfing of PINs) as well as of acquiring it using electronic tools (like biometric skimmers and key loggers). Second, we allow agents to tamper with software on their devices; so, for example, agent A can modify a banking application on \( D_A \) and cause it to replace messages of the form \((w, ID_C, ID_A, y)\) with \((w, ID_C, ID_A, y + \Delta)\) while still displaying legitimate confirmations on the display. Eavesdropping and tampering attacks seem more feasible in branchless banking than in ATM-based banking due to the proximal presence of agents in every transaction.

We assume similar capabilities for customers as well: they can eavesdrop on inputs of any agent they transact with and can modify applications on their respective devices. Finally, we allow all clients in the system the power to acquire another client’s device and to tamper with it or to execute banking protocols using it. This models the possibility of device theft—a security concern in any mobile-based application.

### 3.5 Security

For security, we require that whenever an agent A and a customer C complete a withdrawal/deposit protocol and at least one of them is honest, then the view of the protocol from the perspective of each honest party must be exactly the same as the view from the perspective of the bank. For example, if C honestly completes a withdrawal protocol (using his own untampered device) and C believes that the protocol has decremented his balance by \( x \) and incremented A’s balance by \( y \), then the bank’s database \( T \) must reflect exactly this result. We do not require privacy of protocol transcripts since this seems non-trivial to ensure in the presence of eavesdropping agents. Future work will present more formal security definitions for branchless banking, cover protocols other than deposits and withdrawals and consider stronger security notions; e.g., transcript privacy and security against collusion attacks are two issues we hope to model later on.

### 4 Towards a Secure System

We present some initial ideas on how branchless banking systems that meet the security criteria of the above threat model could be designed. Our description is sketchy at this point; filling in all details is part of our ongoing work.

#### 4.1 Assumptions

We assume all client devices are GSM-compatible mobile phones. Furthermore, we require that some of these phones are programmable and capable of transporting information generated using cryptographic schemes; e.g., these phones might support J2ME and have a GPRS interface. We consider two scenarios.

**Scenario 1: All phones are programmable.** This is the more restrictive scenario since non-programmable phones dominate the mobile phone market in developing regions [17]. Still, we consider it since it simplifies system design considerably and requires no further assumptions. We remark that there is ongoing work on extending the capabilities of non-programmable phones using cheap hardware accessories [8]; such accessories, if produced at scale, can greatly increase the feasibility of Scenario 1.

**Scenario 2: Agent phones are programmable.** This scenario is the more achievable one. Since agents are typically better off than customers and are even compensated for their work, they are more likely to either own a programmable phone already or be convinced into acquiring one. Some banks may even be in a position to commission a programmable phone to each agent—the phone incurs a one-time setup cost which gets amortized across customers the agent transacts with. Not surprisingly, there are some systems which operate under this scenario [1].

We do not consider a scenario in which none of the phones are assumed programmable i.e., where all phones are capable of only basic messaging protocols like SMS, USSD and voice. Although this is the most practicable scenario and is assumed by almost all existing systems, it is difficult to design fully secure systems in this scenario, given recent attacks on GSM telephony [18].

#### 4.2 The Choice of User Credentials

Since our model enables clients to eavesdrop on each other’s inputs during transactions, the choice of credentials to use for authenticating clients is critical for secure system design. Passwords are a weak candidate authentication tool since once an eavesdropper learns a client’s password he or she can trivially impersonate as that client and conduct fraud. Biometrics are less easy to forge and have significant usability advantages which is why several banks, including branchless ones [1], do use them in practice. However, we argue that biometrics, too, are an inappropriate authentication tool in branchless banking. First, voice biometrics are unsuitable because of the ease with which they can be duplicated and the difficulty of deploying them in the real world; deployment is particularly a challenge for developing regions where the incidence of noise pollution is high. Second, biometrics like fingerprints and iris scans are presently expensive to implement, expensive enough that using bio-scanners on a per client basis (the most secure option) is infeasible. Third, if one deploys a single bio-scanner per agent, as done by [1], the cost of deployment may become bearable but the security risks escalate: agents gain immediate access to biometrics of all customers they transact with and this, along with the
threat of software tampering, makes fraudulent use of customer biometrics easier than in the one-scanner-per-client setting.

We also note that by their very nature, biometrics are an irrevocable form of credential; there is a limit to the number of times a credential can be replaced once compromised. Thus, unless we assume complete honesty of all agents in our system or tamper-proofness of scanning software (both difficult possibilities!), biometric-based authentication is a risky proposition in our application.

We propose one-time passwords (OTPs) as the primary user authentication tool in branchless banking. At the time of enrolment, every client \( X \) would receive a list of OTPs \( \langle s_1^X, \ldots, s_n^X \rangle \), generated randomly and independently, for some large \( n \) and a copy of this list would also be maintained at the bank server. The \( ith \ OTP \ s_i^X \ would be used for authenticating \( X \) in the \( ith \) transaction initiated by \( X \). Since each OTP is meant to be used only once, the threat from eavesdropping attacks would be drastically reduced. In implementations, OTPs could be stored on small electronic dongles with small displays [4] or in paper form. As is standard practice, 6-digit OTPs should provide sufficient security.

We remark that a branchless bank named Eko [2], with a client base of over 300,000, has already deployed OTP tokens for user authentication in practice. A security weakness in Eko’s scheme and a potential fix are reported in [14]. (The fix was devised in joint work with Eko.) Even with the fix, the system seems vulnerable to spoofing attacks as launched against M-Pesa [16], although no real execution of such attacks is known.

4.3 Preventing Spoofing and Tampering

To address the issue of spoofing and tampering attacks, we recommend the use of a standard cryptographic tool—digital signatures. Signatures are easier to deploy in scenario 1, so we consider that scenario first. One candidate solution would be to use a public-private key pair for the bank, say \( \langle pk_B, sk_B \rangle \), and one for each agent in the system. All phones would be equipped with the public key of the bank \( pk_B \) during enrolment; in addition, the phone of every agent \( A \) would store the public-private key pair of \( A \), \( \langle pk_A, sk_A \rangle \), which would also be generated by the bank when enrolling \( A \). Client-originating transaction messages would be sent through agent phones only; e.g., for a withdrawal transaction that customer \( C \) conducts with agent \( A \), the former would enter all details, including credentials, into \( D_A \) and the device would sign the information under \( sk_A \) before transmitting it. Since \( A \) could potentially tamper with her phone’s software, it is critical that the bank’s confirmation messages are also signed and later verified by each client device; e.g., a confirmation message for a withdrawal request made by \( C \) must be signed under \( sk_B \) and \( D_C \) must verify the signature using \( pk_B \) before approving the message. In order to prevent messages from being replayed, a timestamp should accompany every message that is transmitted.

4.4 Addressing Scenario 2

Scenario 2 is more challenging, and arguably the more interesting one, to tackle. In this scenario, customer phones cannot be programmed, which essentially means we need techniques other than digital signatures to authenticate messages sent by the bank. While it may not be possible to achieve foolproof security under such a constraint, our belief is that we could get pretty close without significantly hampering efficiency.

Relaxing the model. One approach would be to consider security in a slightly-relaxed threat model wherein we rule out attacks which are known to be relatively improbable in practice. For example, in our current model, the assumption that adversaries can spoof messages on behalf of other parties is quite realistic, but that they can launch man-in-the-middle (MITM) attacks arbitrarily is less so. While GSM’s design makes MITM attacks feasible for sessions originating from a mobile phone, doing the same when only the recipient is mobile, but the session originator is not, is still relatively difficult [18, 12]. If the bank negotiates a reliable SMS channel with its network operator for push messages and a secure TCP/IP link to connect to the operator’s SMSC, the risk of MITM tampering on the messages it sends to customers is drastically reduced.

If we rule out MITM tampering on bank-originating messages, we mainly need to guard customers from message spoofing attacks. One way to do this would be the following: The bank provides to each customer \( C \) a list of response OTPs \( \langle r_1^C, \ldots, r_r^C \rangle \), generated randomly and independently of each other and of the \( s_i \)’s that \( C \) already received. For every confirmation message targeted at \( C \), the bank appends the first unused OTP in this list to the rest of the message and sends it via SMS to \( D_C \). \( C \) verifies authenticity of the SMS by checking that it ends in the first unused response OTP on his list. If this is not so, the indication is that the SMS has been spoofed. Alternate mechanisms to send confirmation messages (e.g., using voice or USSD) could result in different grades of security and usability; understanding these trade-offs thoroughly would require further research.

Special-purpose signatures. A more radical approach would be to design new signature schemes for which signature verification is computationally cheap and can be directly performed by humans. Although we believe it is impossible to design humanly-verifiable digital signatures which provide reasonable security and work for arbitrary messages, in the context of branchless banking, where messages have limited variability, a satisfactory solution may very well exist. The main variables in a bank’s con-
firmation message which need verification are the transaction amount and the transacting agent’s identity and these can be represented with about 10 digits in most implementations. Furthermore, for a given customer, the number of possible values of these variables is often small enough to make them representable with even fewer digits. In ongoing work, we are exploring the design of symmetric-key signatures (i.e., MACs) which operate on short messages and are verifiable by humans with basic numeracy skills.

4.5 On 2-factor Authentication

One limitation of OTP-based systems is that they rely on the use of physical tokens, which are susceptible to loss and theft. For this reason, it is generally recommended that in user authentication tasks, OTPs not be used solitarily, but as a supplement to other tools like passwords or biometrics. As discussed earlier, it is risky to deploy the latter tools in branchless banking systems due to the eavesdropping power available to agents. We make the following suggestion: use numeric passwords (PINs) along with OTPs but leverage human computation capabilities to combine OTPs with PINs in a way that foils eavesdropping attacks against the latter. One candidate solution is the substitution-based coding technique of [14], which enables users to mentally transform 4-digit PINs into random 4-digit numbers using 10-digit one-time keys. The transformation scheme of [14] has been shown to be usable by low-literate users and is also a secure 2-factor solution; indeed, Eko, which was involved in developing the solution, plans to deploy it in the near future.

5 Conclusion

Despite being celebrated as one of the most promising solutions to financial inclusion in the world, today’s branchless banking systems seem to fall short of providing good security guarantees to their beneficiaries. While some may view security of such systems with lesser importance than other design criteria, we believe that security issues are particularly a concern here because of the massive cash flows these systems tend to generate and, at the same time, the limited educational background and negotiating power of the users they serve.

This paper begins the process of understanding and resolving the issue of end-to-end security in branchless banking systems. We present the first security model for this problem and are currently working towards designing the first solution which addresses not only the requirements of this model but also the operational constraints under which the solution is expected to be used. Our hope is to arrive at an open set of standards for developing end-to-end secure and deployable branchless banking systems and to improve current practice based on these standards.

6 Acknowledgements

Thanks to Bill Thies, M. Satyanarayanan and the reviewers of HotMobile 2011 for their valuable feedback.

References