Challenges and Architectural Approaches for Authenticating Mobile Users

João Pedro Sousa
George Mason University
Computer Science Department
4400 University Drive 4A4, Fairfax VA
+1-703-993-9196
jpsousa@cs.gmu.edu

ABSTRACT
This paper casts an architectural eye at existing work on security and privacy in mobile computing. Specifically, it focuses on authentication as it leads up to access control from two points of view: servicer providers granting access to users, and users granting access to service providers. The paper identifies three classes of problems addressed by existing work. Then, it teases out architectural patterns being used to address those problems, their assumptions, properties, and remaining challenges.

Categories and Subject Descriptors

General Terms
Design, security.

Keywords
Access control, anonymity, authentication, mobile computing, trust, ubiquitous computing, user mobility.

1. INTRODUCTION
Emerging mobile and ubiquitous computing technologies raise frequent concerns about the security and privacy implications to their users. Although a variety of work has addressed mobile computing security, it is not always clear what problem is really being solved and what assumptions are being made.

Questions include: what new problems does mobile computing raise that cannot be addressed by classical solutions, such as Kerberos? Can work in zero-knowledge proofs and peer-to-peer authentication help address some of the new problems? Are anonymous authentication mechanisms a panacea? For what kinds of problems are they not appropriate? Is there a tradeoff between security and ease of access to services available in smart spaces?

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SAM’08, May 10, 2008, Leipzig, Germany.
Copyright 2008 ACM 978-1-60558-022-7/08/05...$5.00.

This paper casts an architectural eye at the problem domain and mechanisms for authentication, leading up to access control from two points of view: servicer providers granting access to users, and users granting access to service providers. In the remainder of this paper, section 2 identifies classes of problems, while section 3 discusses two fundamental pieces for building solutions: trust, and its tradeoffs, and different kinds of credentials, and their strengths and weaknesses. Sections 4 to 7 focus on different architectural patterns that can be recognized in current work, relate those patterns to the problems being addressed, and identify the assumptions, properties, and challenges associated with such patterns. Section 8 concludes and identifies challenges still largely unaddressed by existing work.

2. PROBLEM DOMAIN
Existing work in authentication for mobile computing addresses three classes of problems:

User Access to Services (UAS). In this class of problems, users are willing to release, and present proof of, their identity in order to access owned resources or personalized services. The user, or a device supporting the user, needs to have prior knowledge of an identifier through which the service provider can be reached. Examples include: a user opening a work session on a local or remote machine (e.g., telnet) and user accessing a website that keeps personalized accounts, such as a bank or travel website. Services that involve electronic payment often require this kind of authentication, although some allow GAS (below).

Group Access to Services (GAS). In this class of problems, users wish to stay anonymous but are willing to either prove their right of access or their trustworthiness, or in some cases, to release profiling information about themselves. The resources or services being accessed are not owned or personalized to individual users, and are generally available to a group of users, either defined by enumeration or by characteristics (e.g., customer segments). Whenever the identity of the user can be reduced to a group of $k$ known users, such as in the case of membership-based access, the user is said to have $k$-anonymity [27]. Three sub-classes of problems can be distinguished, according to the criteria used to admit anonymous users: right to access (e.g. membership-based access and electronic voting,) trustworthiness (protection against malicious users,) or improved user experience (e.g. e-commerce and targeted advertising). In either of these cases, the providers may be known beforehand, or may announce their availability and be discovered opportunistically by users.
Electronic payment may be supported by certified electronic currency, which users must obtain from trusted third parties (to whom they authenticate in the UAS form,) and which providers redeem at the same third parties. Service providers may also choose to require a user to upgrade the authentication to UAS when it comes to payment, in order to obtain direct traceability.

**Link Peers (LP).** This class of problems focuses on establishing a secure link between devices. The identity and willingness of the peers is normally established by external mechanisms; for example, two users who decide to link their personal devices to exchange information, or a single user that decides to pair devices for supporting tasks such as remote control, or media streaming.

Real situations may require solving problems in more than one class to achieve the user’s goals. For example, suppose that a user wants to access a web-based media library using (a) a wall-mounted display at a lounge and (b) the user’s mobile phone as a remote control to playing the media. Three problems need to be solved: a GAS, to certify the user’s membership and access to the media library; and two LPs, link the user’s phone to the specific display close to the user, and link that display to the media server.

### 3. SOLUTION BUILDING BLOCKS

This section discusses the roles of trust tradeoffs, as conceptual framework, and of presenting credentials, as mechanism for building solutions.

The notion of trust plays a crucial role in using computer systems. For example, when a user downloads a piece of code, such as an applet, being able to verify the code’s certificate does not prove that the code itself is safe. It proves that the code was issued by a particular publisher and that it was not tampered with while in transit through the network. If a user decides to trust the code, that decision is based on his or her appraisal of the publisher’s reputation.

Everyday life is full of situations where privacy, security and functionality are at odds, and trust plays a key role in managing such tradeoffs. For example, when someone on vacation decides to use an automated teller machine in a store, or to enter an internet cafe and logon to an online banking system, the person is ultimately risking having his or her authentication credentials stolen as the result of malicious or unsafe infrastructure. However, people assess the trustworthiness of their surroundings. Someone may decide not to enter a restaurant in a bad neighborhood, or may decide to have only well cooked food if unsatisfied with the general appearance and cleanliness.

An important insight is that trust is both purposeful and selective. For example, a user may deem the computing environment at a cafe to be unsafe to carry out online financial transactions, but quite acceptable for sharing online vacation photos with a friend. This is in contrast with pursuing an absolute view of trustworthiness, where a user or a service is either trustworthy or not. This latter view is sometimes adopted in more traditional applications, where trust evaluation is taken in the restricted context of that application. However, in the broader context of mobile and ubiquitous computing, it is important to keep in mind that trustworthiness must be framed by the user’s goals.

Fundamental for building trust among parties is certifying that the parties are who they say they are. Presenting credentials is the base mechanism to authenticate a principal, as the preamble for granting access to the desired resources or services. A principal may either be a user or a device, and what is authenticated may either be the principal’s identity, or its right of access/trustworthiness, while keeping the principal anonymous.

A wide range of credentials can be presented for authentication. Different kinds of credentials have different strengths and weaknesses, such as the cost of being presented, or being subject to forgery or false identification. To circumvent these limitations, credentials may be combined to complement each other’s weaknesses. Specifically, credentials are normally classified into:

**What you know** for example passwords, which are easy to change and easy to keep private, but can be hard to keep track of [2, 13]. These can also be disruptive to provide, as Bardram observed in the case of medical staff working at a hospital [5]. At the high-end of technology to provide, zero-knowledge proofs allow a principal to prove that it knows a secret, without revealing either the secret itself, or anything else about the principal’s identity [6]. These proofs involve several rounds where the challenger asks the principal different questions about a computationally hard problem, to which the principal must convincingly demonstrate it knows the solution.

**Who you are** for example fingerprints, face and gait recognition [23]. These are hard to change and hard to keep private: e.g., a user’s gait can be lifted by a malicious hall without the user’s knowledge. Moreover, some of these are susceptible to false positives [19]. These are however, very easy to provide: the user just needs to be him or herself.

**What’s in your vicinity** which is the generalization of “what you carry” and “where you are.” For example the user may carry smart cards or one-time password devices [29]. Also, for example, if the user is standing next to a device inside a secure room, that means the user had the means to enter the room in the first place. These credentials may preserve anonymity (e.g., someone used a key to enter the room) are feasible to change (make a new set of keys to the secure room and distribute to everyone who should have access,) feasible to keep private (don’t loose the keys,) and somewhat easy to use (remember to bring the keys.)

### 4. TRADITIONAL UAS

Traditional authentication verifies the identity of the user and uses that identity to establish the set of resources to which the user has access. Although simple solutions exist where the service provider keeps knowledge of the user’s credentials, modern solutions, such as Kerberos [15], use a trusted third party to issue tickets valid to a specific server. Such tickets, however, include the identity of the requesting user, so that the provider can derive the corresponding Access Control List (ACL).

Figure 1 shows an informal architectural diagram of traditional authentication, where connectors A and T typically coordinate their behavior to implement some variant of the Needham-Schroeder authentication protocol [21]. Initially, the user approaches a workstation, which may be shared, and reveals his identity and credential (typically a password) as well as the identity of the desired server.

Traditionally, connector T sends the user id in the clear (but not the password), and then both ends use their local knowledge of the password (the ticket issuer stores it, as result of user registration) to generate a symmetric key for securing further
communication in T. In more advanced schemas that use a public key infrastructure (e.g., [22]), the WS may open an anonymous secure link to the ticket issuer using the latter’s public key. In either case, the communication across A is secured by a symmetric key generated by the ticket issuer.

This pattern for authentication is by far the most used today, due possibly to the prevalence of client-server architectures. The pattern assumes strong connectivity to both the server and to the tickets issuer, which becomes a limitation in mobile environments with weak connectivity.

From a security standpoint, this pattern is characterized by the leap of faith required from the client (user,) who has to reveal his identity and intention to communicate with a specific server across the wire to a third party, before the identity of that party itself has been authenticated. Furthermore, the user implicitly needs to trust the server: this pattern is designed for the protection of service providers against malicious users.

5. LINKING PEER DEVICES

With the increasing popularity of mobile devices and of peer-to-peer sharing applications, mutual authentication of peer devices is an increasingly relevant problem.

Figure 2 shows two informal architectural diagrams, where the rounded corner rectangle with the dashed line represents user ownership of the devices inside the rectangle.

For the case shown on the right of Figure 2, the problem is to securely link devices belonging to the same user and within a short range of each other. While work such as Stajano and Anderson’s follows an approach where knowledge about peers and their roles are loaded into devices [26], which is similar to the multi-user case discussed below, other work requires no prior knowledge of the peers.

Mayrhofer and Gellerson use shared accelerometer data as the base to securely link two devices with no prior knowledge of each other [20]. The two devices are shaken together, which enables them to independently come up with the same encryption key. This encryption key is used on a short-range broadcast, connector S, which can only be deciphered by the shakenly joined peer.

For the case shown on the left of Figure 2, where multiple users carry mobile devices, a body of work exists. Abadi et al. propose an authentication scheme that hides the identity and location of two communicating principals from third parties [1]. Also, work in ad-hoc access control addresses scenarios where two users with personal mobile devices want to share content, establishing access control over ad-hoc networking either without the need of a central authority [3], or relying on a local coordinator which disseminates a membership list establishing role-based access control [14].

In contrast with local networks or short range networks, such as Bluetooth [9], securing connector R over wide-area networks is a more complex problem. Rannenberg et al. studied multilateral security in the telecom domain, where users, service providers, and carriers may each have their own, possibly conflicting, security and privacy concerns and policies [24]. Rather than imposing the “right” solution, Rannenberg allowed users to tune their security solutions, and showed that users at different stages of interest, understanding, and competence learned to understand the consequences of their decisions and tuned their policies to reach a satisfying privacy level.

In this pattern, peers need to learn about each other by some external mechanisms. As a result, peers also store identifiers for reaching each other (such as URLs, Bluetooth ids, or phone numbers,) in addition to security information, such as the peers’ public keys. The authentication of the users to the devices, and therefore among different users, is implied by device ownership: a what you carry credential, as described in section 3.

6. GAS IN SHARED SPACES

A traditional pattern, such as depicted in Figure 1, could support anonymous access to services if the user reveals a group identifier and credential to the WS. However, a more interesting case is when a user carrying a mobile device wishes to access ambient services at shared smart spaces; for example, browsing internet news at a lounge, following an IT-supported visit to a museum, or inquiring for nearby restaurants and bus schedules at a bus stop.

Figure 3 shows a pattern to address this latter case, where device ownership is represented as before, and a connector going through the box for a component means that the realization of that connector requires the willingness and support of that component or its substructure. Specifically, connector T, with similar purpose and properties as discussed in Figure 1, depends on the willingness of the ambient services to relay the communication. In real situations, several organizations may contribute to the ambient services available to a user at a particular location, for example, the PDA may see several overlapping wireless network cells in addition to mobile phone networks.

The pattern in Figure 3 supports the three sub-classes of problem identified in section 2, and which report to the criteria used to admit anonymous users: right to access, trustworthiness, or improved user experience.

In all sub-classes the ambient services associate group identifiers, as opposed to user identities, to the corresponding access control or service level policies. The user’s identity is known only to the
PDA and, in the case of right-to-access, also to the certificates issuer. Typically, ambient services are willing to announce their presence and identity.

Right to access. Proposed solutions have exploited the properties of mathematical functions and encryption to produce one-time unique credentials (e.g. Chowdhury et al [8]). The user’s PDA opens a secure connector T to a certificate issuing authority, e.g. using public key encryption, and proves the user’s identity to the authority. The latter then issues an anonymous certificate asserting that the user has the right to access the requested service. A different certificate is produced in this way every time the user wishes to access a service, which prevents the service provider from tracing the identity or usage patterns of individual users.

However, the identity of the user may be revealed to the service provider, by means of connector V, if the latter can justify to the authority that the user abused the service in any way. A further threat to the user’s privacy is the fact that the authority can trace the usage patterns of individual users.

To reduce the exposure of usage patterns to the issuing authority, Teranishi et al. proposed that the authority issues k anonymous certificates (k depending on the nature of the problem) which then can be used at discretion [28]. The properties of such certificates are such, that if a malicious user tries to use the service k+1 times, by reusing one of the certificates, the service provider can compare the two usages and recover a base of the certificate that it then shows to the authority in order to identify the malicious user.

Zero-knowledge proofs are excellent candidates to address this problem, since they can certify a user’s right to access, while not involving a third party, and not revealing the identity of the user to the service provider [6]. Currently their use is limited because of complexity and computational overhead.

Trustworthiness. To guard against the actions of malicious software, a group of work builds on the Trusted Computing (TC) infrastructure [4]. In this case, group membership is defined by trustworthiness. The certificates issuer in the TC infrastructure is called Direct Anonymous Attestation Issuer and engages in protocols over connectors T and V for the mutual verification of trustworthiness. These protocols relies on software running on each device that verifies the integrity of the entire platform and produces signed certificates thereof (e.g., [12, 17]).

User experience. This sub-class of problems occurs in public spaces such as shopping malls and museums, where the goal of the ambient services is to better serve segments of users with different interests. For that, the ambient services may try to learn the user’s demographics and personal interests, or they may want to keep track of usage patterns of regular users. Users, on the other hand, may want to preserve their privacy, revealing as little as possible about themselves.

A frequent approach is for the PDA to release a pseudonym for the user, to prevent further data to be associated with the real user identity. For example, Chatfield and Hexel propose that the user’s PDA stores a set of pseudonyms for the user, and manages the release of the appropriate pseudonym at each location, with the assistance of location awareness (e.g., [10]). The ambient services store information such as usage history and preferences associated with each known pseudonym.

However, storing information for each pseudonym with the ambient services has a number of limitations. First, scalability on the number of users, for spaces with large numbers of infrequent users, such as airports or museums. Second, inability to reuse information across spaces of the same type: for example, a user is likely to have similar preferences in all airports. And third, the learning curve that results from each new space having to learn user preferences from scratch.

An alternative approach is for the PDA to store user preferences and usages patterns associated to each type of space, or to each type of user activity (e.g., [25]), and to release that information under a different pseudonym each time. Additionally, the user may allow some profiling information to be released associated to each kind of activity, so to enable ambient services to propose alternative options that the PDA has not encountered before.

7. UAS IN SHARED SPACES

A more ambitious goal is to allow mobile users to leverage the capabilities of ambient services for accessing remote personal resources and services. For example, sharing vacation pictures with family at home and later with a friend at a coffee shop; working on a document at the office and later discussing it at a meeting; or accessing a patient’s medical information all the while the patient moves among different units at a medical facility.

Figure 4 shows a pattern that combines access to ambient services, E, with access to personal resources and services, H, and the user is carrying a personal mobile device, such as a PDA. In the following discussion, we will refer to the PDA as a representative of the broad class of personal mobile devices, including smart phones and laptops.

The prevalent approach today is for users to separate their access to ambient services (as is section 6) from their access to personal resources (as in section 4, where the WS is replaced by the PDA).

Despite its similarity to traditional authentication, the use of a PDA, as opposed to a professionally maintained WS sitting at an office or lab, brings new challenges. To make accessing H easier, users may be tempted to cache their credentials (e.g., in the form of saved passwords in browsers) and H’s identifiers in the PDA.

Unfortunately, users often neglect to secure access into personal devices, e.g., not defining locking passwords on their cell phones, under the assumption that nothing bad will ever happen. Despite
the sense of security given by ownership, it may also lead to severe compromises in case the personal devices are lost or stolen. Users may also not have the skill or the will to protect their mobile devices from cyber attacks. Unfortunately, if the PDA is infiltrated, it can be used as an entry point to all personal resources and services.

These challenges can be addressed in part by improving H’s awareness of the potential risks, for example, associated to each location the user is at. Lee et al. address the problem of dynamically adjusting access rights H grants to the PDA depending on its location [16]. Hoffmann et al. proposed an infrastructure of mobile agents to support multilateral security in a scenario divided in two domains: the internet, and the user’s security domain, which contains one or more mobile devices and a home base, an always-on agent server providing access to security policies and profiles [11].

An issue that is not addressed by these solutions is if the PDA is compromised at a high-risk location, for which H granted low access rights to the remote resources, but the malicious software hides and only manifests itself at a location that H deems low risk.

Furthermore, the approach of separating UAS from the ambient is limiting, or optimist, since it assumes that users will not try use ambient capabilities, such as large displays, to carry out activities that involve personal resources and services.

If E is willing to cooperate, it is still possible to keep the identity and credentials of the user known only to the PDA and to H: the PDA can mediate the access to H, which then pushes the relevant information onto E. Trust between E and H can be established anonymously by a trusted third party [4, 8].

7.1 Same, with no PDA
Relying on a PDA to always mediate access to H may become cumbersome. For example, medical staff working at a hospital move from room to room and need to access patient information at the most convenient devices. Remembering to carry the PDA, and making sure batteries don’t run out may become an unacceptable overhead. Even in more informal situations, users may wish to access personal resources at a shared smart space, regardless of not carrying a PDA.

Figure 5 shows a pattern that combines access to ambient services, E, with access to personal resources and services, H, but the user does not carry a personal mobile device. In this case, the user identity and credentials, as well as an identifier of H, must be revealed to E so that E mediates access to H, and H grants access to the user’s personal resources. This form of authentication can be achieved with little or no distraction to users by leveraging credentials in the who you are category (see section 3) or simply smart badges [5].

With respect to threats, if E is malicious, it can capture the user’s credentials and replay them later to gain access to the user’s information. Possible countermeasures for replay attacks are to employ one-time session credentials [29]. A malicious E can also piggyback illicit requests while the user is in session and without the user’s awareness, something one-time credentials provide no protection against.

As a countermeasure, H can keep a log of remote accesses by E, which can be leveraged for accountability purposes.

Even if no attack is directed towards H, a malicious E can reveal the user’s presence and context to third parties, for instance whether the user has company. Celebrities are likely targets of this kind of attack.

If E is not malicious but not secure enough, a malicious third party may infiltrate E and perform either of the attacks above without E’s awareness.

On the other hand, devices embedded in the premises of reputable and accountable entities are likely to be professionally administered and harder to compromise than a mobile personal device. In other words, the pattern in Figure 5 by no means should be outright dismissed as worse than the one in Figure 4.

Despite the difficulty to provide foolproof solutions, it is possible to manage the risk associated with this pattern. In essence, this becomes a problem of controlling access: determining the access rights that the user is willing to grant E, depending on the user’s needs and trust on E (see section 3).

A body of work concerns the evaluation of trust. For example, Wu et al. apply federated trust management to pervasive healthcare systems [30]. Fuzzy logic is used to handle objective (can be measured) vs. subjective trust. Beth et al. categorize trust relationships into two classes, direct trust and recommended trust, to come up with a numeric value for trust based on the past experiences of different actors [7]. Manchala explores the notion of quantifiable trust for electronic commerce. Formally, trust is characterized as a binary relation between consumer and provider, from in the context of a specific service or application [18].

However, while trust quantification can offer an abstract assessment of risk, trust is often associated to a purpose, rather than to an abstract quantitative scale. For example, a user may trust his doctor’s smart office for accessing his medical records, and may trust his financial consultant’s smart office to access his retirement plan. In examples such as this, rather than thinking in terms of which purpose requires a higher level of trust, it seems more natural to describe what purpose each location is trusted for.

Frameworks to manage access based on a notion of purposeful trust, to the author’s best knowledge, are currently missing.

8. CONCLUSION
This paper reviewed a range of work dealing with authentication, and classified that work according to the class of problem it addresses and to the architectural pattern it adopts. The paper also discusses assumptions and properties of each pattern.

An outstanding issue is that current work offers point solutions for a single problem or application.

Missing are integrated frameworks for authentication that can address situations such as the example mentioned in section 2, where a user wants to access a web-based media library using (a)
a wall-mounted display at a lounge and (b) the user’s mobile phone as a remote control to playing the media. This relatively simple situation requires solving three authentication problems in two distinct classes, also identified in section 2.

A challenge for such frameworks is to help users manage their credentials and their release in a manner that reconciles the user’s goals for accessing functionality with the user’s privacy and security preferences. Specifically, not revealing more about the user than necessary, and advising the user about possible leaks or threats associated with a particular course of action.

Future work on such frameworks should clarify the role of infrastructure, such as trusted third parties, versus the role of personal mobile devices. And, ideally, come up with a model that works, although possibly in degraded modes, regardless of the availability of personal devices and of the reachability of trusted third parties.

9. REFERENCES