US 20230153464A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2023/0153464 A1**

Boffa et al. (43) **Pub. Date: May 18, 2023**

(54) **SYSTEMS AND METHODS FOR SECURE DATA SHARING**

(71) Applicant: **YALE UNIVERSITY**, New Haven, CT (US)

(72) Inventors: **Daniel Boffa**, Branford, CT (US); **Michael Fischer**, Hamden, CT (US); **Jonathan Hochman**, West Hartford, CT (US)

(21) Appl. No.: **18/057,083**

(22) Filed: **Nov. 18, 2022**

Related U.S. Application Data

(60) Provisional application No. 63/280,955, filed on Nov. 18, 2021.

Publication Classification

(51) **Int. Cl.**

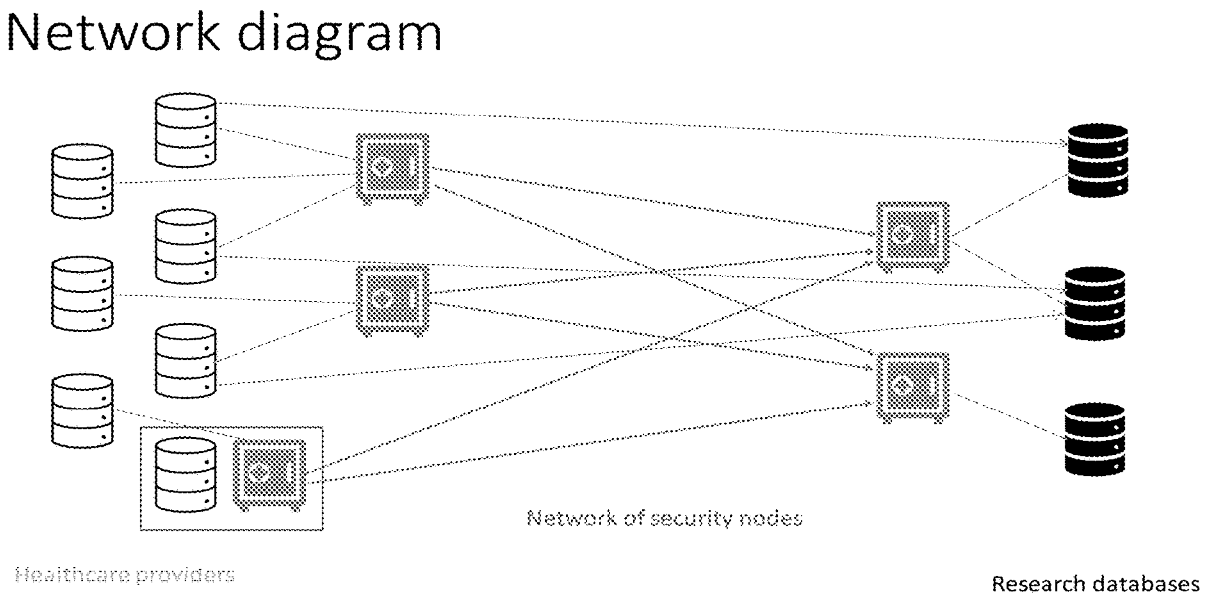
G06F 21/62	(2006.01)
G06F 21/60	(2006.01)
H04L 9/06	(2006.01)

(52) **U.S. Cl.**

CPC ***G06F 21/6254*** (2013.01); ***G06F 21/602*** (2013.01); ***G06F 21/6227*** (2013.01); ***H04L 9/0643*** (2013.01)

(57) **ABSTRACT**

One aspect of the invention provides a method for secure sharing of data. The method includes: receiving, from a first computing device and by a security node for the first computing device, a hashed identifier for a data source; generating, in response to the receiving, a blinding function value dependent on the hashed identifier; and transmitting, to the first computing device, the blinding function value for storage of a set of data and linking the set of data to the data source.



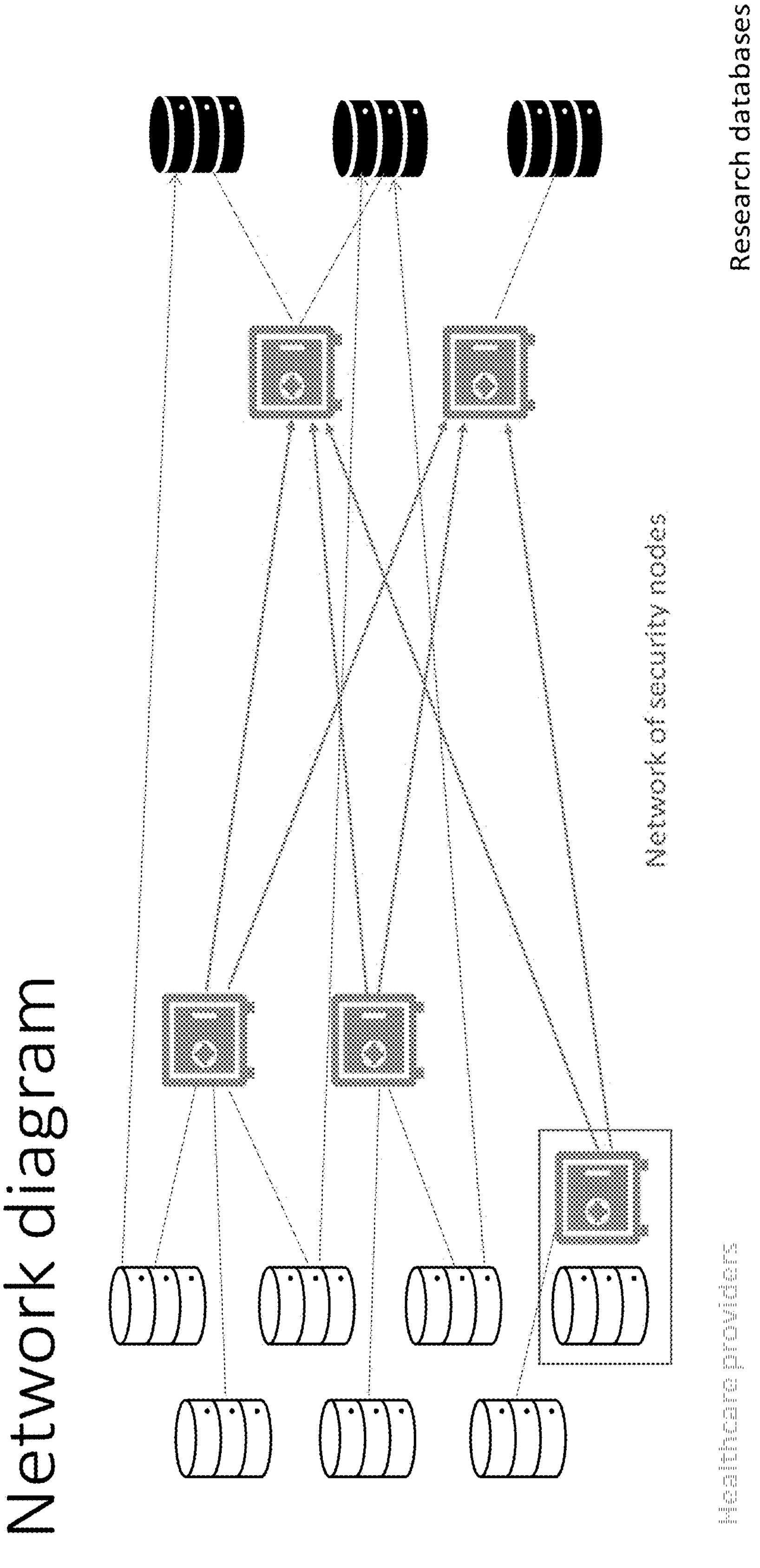

Network diagram
Research databases
Network of security nodes
Healthcare providers

FIG. 1

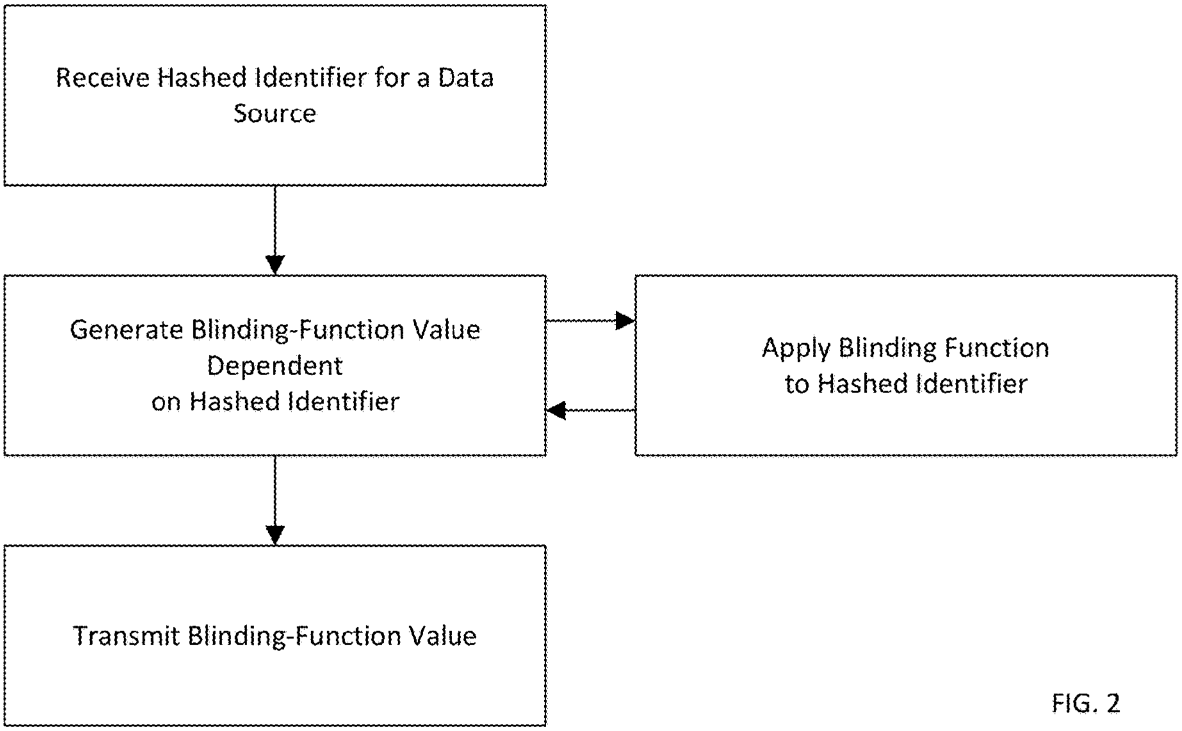

Receive Hashed Identifier for a Data Source
Generate Blinding-Function Value Dependent on Hashed Identifier
Apply Blinding Function to Hashed Identifier
Transmit Blinding-Function Value

FIG. 2

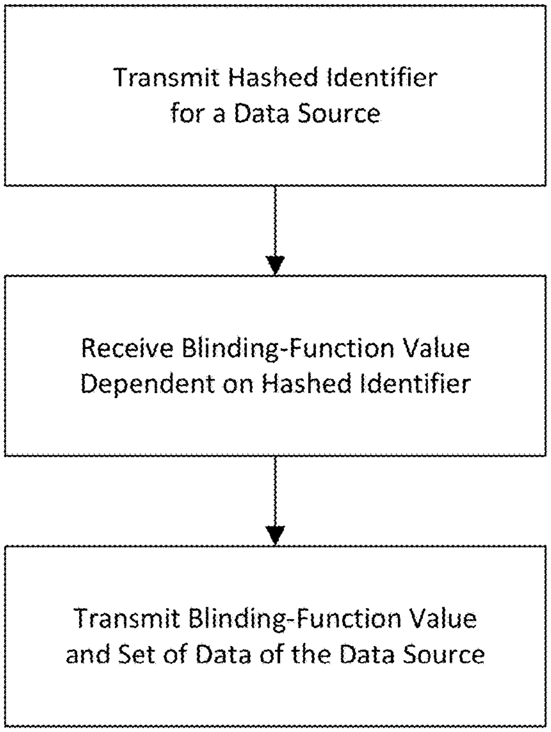

Transmit Hashed Identifier for a Data Source
Receive Blinding-Function Value Dependent on Hashed Identifier
Transmit Blinding-Function Value and Set of Data of the Data Source

FIG. 3

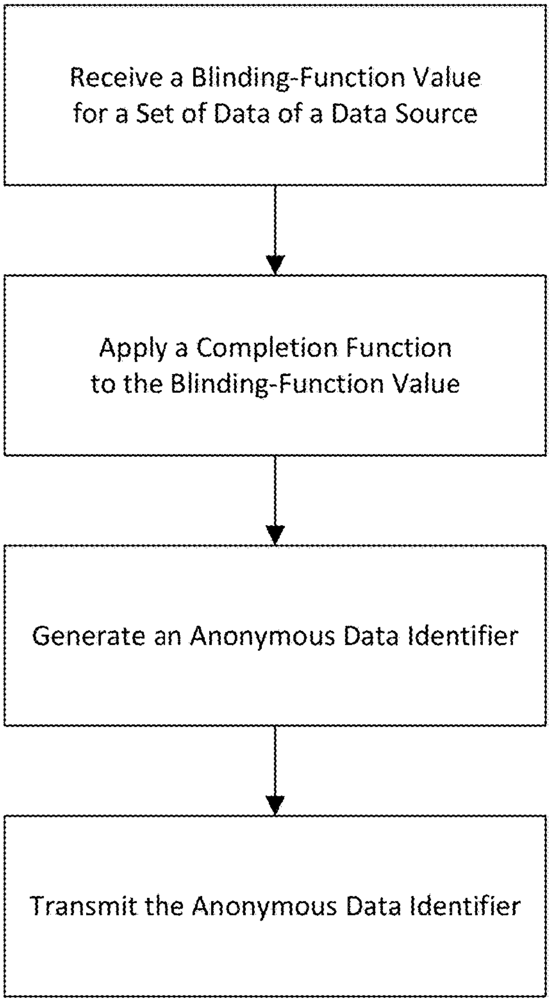

Receive a Blinding-Function Value for a Set of Data of a Data Source
Apply a Completion Function to the Blinding-Function Value
Generate an Anonymous Data Identifier
Transmit the Anonymous Data Identifier

FIG. 4

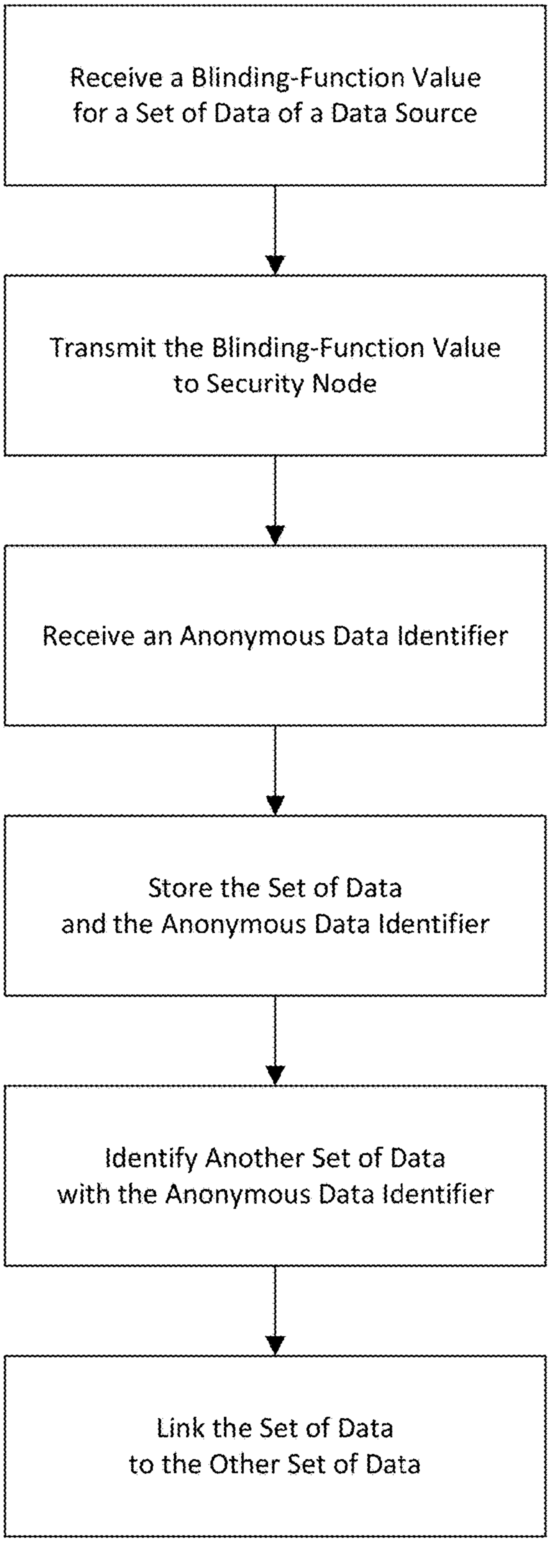

Receive a Blinding-Function Value for a Set of Data of a Data Source
Transmit the Blinding-Function Value to Security Node
Receive an Anonymous Data Identifier
Store the Set of Data and the Anonymous Data Identifier
Identify Another Set of Data with the Anonymous Data Identifier
Link the Set of Data to the Other Set of Data

FIG. 5

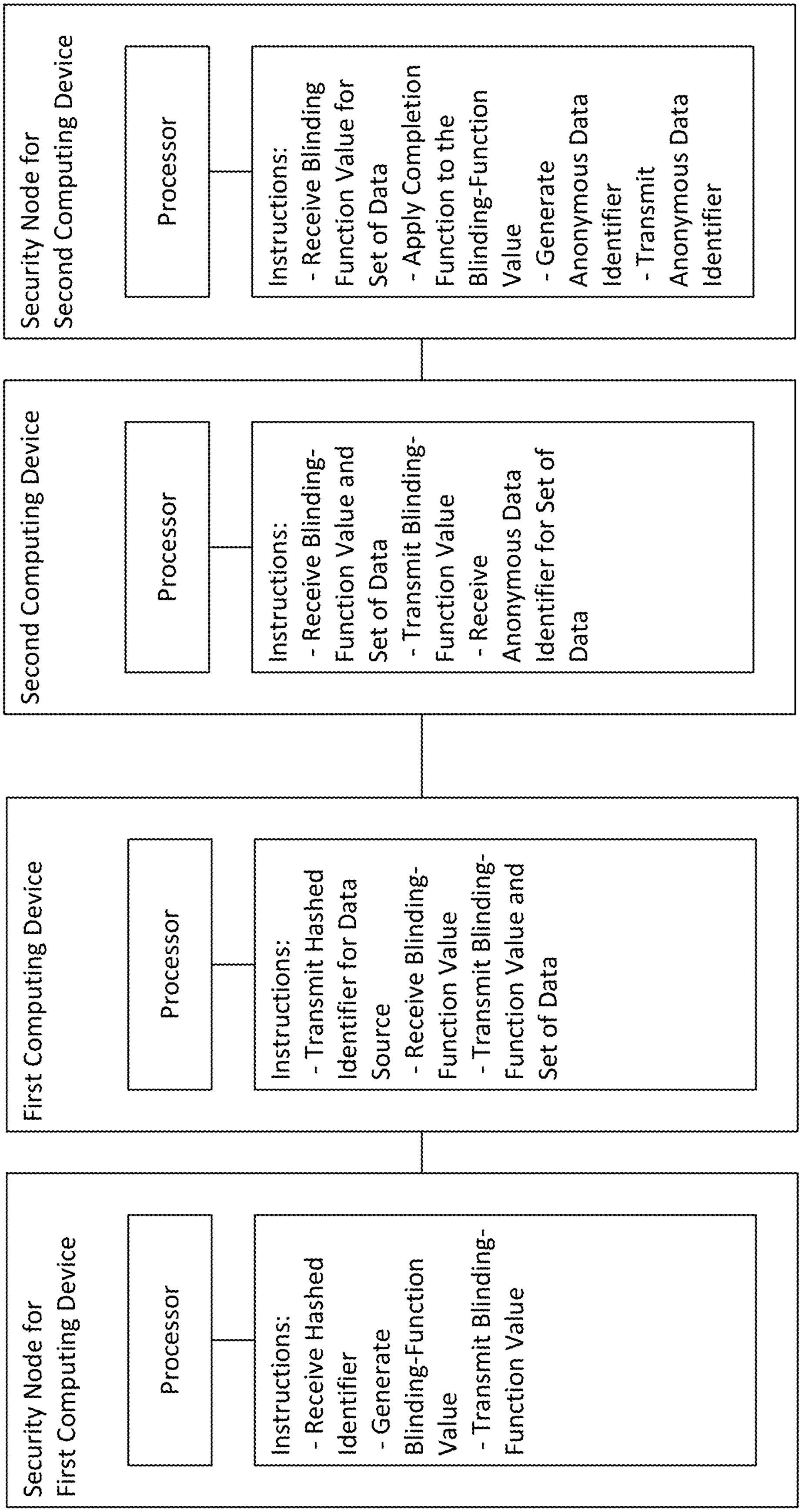

Security Node for First Computing Device
Processor
Instructions:
- Receive Hashed Identifier
- Generate Blinding-Function Value
- Transmit Blinding-Function Value
First Computing Device
Processor
Instructions:
- Transmit Hashed Identifier for Data Source
- Receive Blinding-Function Value
- Transmit Blinding-Function Value and Set of Data
Second Computing Device
Processor
Instructions:
- Receive Blinding-Function Value and Set of Data
- Transmit Blinding-Function Value
- Receive Anonymous Data Identifier for Set of Data
Security Node for Second Computing Device
Processor
Instructions:
- Receive Blinding Function Value for Set of Data
- Apply Completion Function to the Blinding-Function Value
- Generate Anonymous Data Identifier
- Transmit Anonymous Data Identifier

FIG. 6

SYSTEMS AND METHODS FOR SECURE DATA SHARING

CROSS-REFERENCE TO RELATED APPLICATION(S)

[0001] This application claims the benefit of priority under 35 U.S.C. § 119(e) to U.S. Provisional Patent Application Ser. No. 63/280,955, filed Nov. 18, 2021. The content of this application is hereby incorporated by reference herein.

BACKGROUND OF THE INVENTION

[0002] Electronic patient medical records contain vast amounts of information of potential value to researchers striving to increase understanding of diseases, treatments, and outcomes. Effective use of such data is limited by privacy and technical concerns.

SUMMARY

[0003] One aspect of the invention provides a method for secure sharing of data. The method includes: receiving, from a first computing device and by a security node for the first computing device, a hashed identifier for a data source; generating, in response to the receiving, a blinding function value dependent on the hashed identifier; and transmitting, to the first computing device, the blinding function value for storage of a set of data and linking the set of data to the data source.

[0004] This aspect of the invention can have a variety of embodiments. The method can further include applying a blinding function to the hashed identifier to generate the blinding function value. The blinding function can be of the form $b_u(x)=xg^u \bmod p$, wherein: x comprises the identifier for the data source; g comprises a primitive root value; u comprises a secret random value selected by the security node for the first computing device; and p comprises a safe prime value.

[0005] The first computing device can include a device for a healthcare provider. The data source can include a medical patient.

[0006] Another aspect of the invention provides a method for secure sharing of data. The method includes: transmitting, from a first computing device and to a security node for the first computing device, a hashed identifier for a data source; receiving, in response to the transmitting, a blinding function value dependent on the hashed identifier; and transmitting, from the first computing device to a second computing device, the blinding function value and a set of data of the data source for storage of the set of data and linking the set of data to the data source.

[0007] This aspect of the invention can have a variety of embodiments.

[0008] The method can further include generating the hashed identifier by applying a hashing function to an identifier of the data source. The hashing function can include a SHA256 hashing function. The hashing function can include a one-way hashing function.

[0009] The second computing device can include a database configured for storing medical records.

[0010] Another aspect of the invention provides a method for secure sharing of data. The method includes: receiving, from a second computing device and at a security node for the second computing device, a blinding function value for a set of data of a data source; applying, by the security node, a completion function to the blinding function value; generating an anonymous data identifier from the applying; and transmitting the anonymous data identifier to the second computing device for storage of the set of data and linking the set of data to the data source.

[0011] This aspect of the invention can have a variety of embodiments. The completion function can be of the form $c_v(y)=yg^v \bmod p$, wherein: y comprises a hashed identifier for the data source; g comprises a primitive root value; v comprises a secret random value selected by the security node for the second computing device; and p comprises a safe prime value.

[0012] Another aspect of the invention provides a method for secure sharing of data. The method includes: receiving, from a first computing device and at a second computing device, a blinding function value and a set of data of a data source; transmitting, form the second computing device to a security node of the second computing device, the blinding function value; receiving, in response to the transmitting, an anonymous data identifier for the set of data; and storing the set of data and the anonymous data identifier at the second computing device.

[0013] This aspect of the invention can have a variety of embodiments. The method can further include: identifying another set of data stored at the second computing device with the anonymous data identifier; and linking the set of data to the other set of data based on the identifying.

[0014] Another aspect of the invention provides a system for secure data sharing. The system includes: (a) a first computing device, (b) a security node for the first computing device, (c) a second computing device, and (d) a security node for the second computing device. The first computing device is configured and adapted to: transmit a hashed identifier for a data source; receive, in response to the transmitting, a blinding function value dependent on the hashed identifier; and transmit the blinding function value and a set of data of the data source for storage of the set of data and linking the set of data to the data source. The security node for the first computing device is configured and adapted to: receive, from a first computing device, the hashed identifier for the data source; generate, in response to the transmitting, the blinding function value dependent on the hashed identifier; and transmit, to the first computing device, the blinding function value. The second computing device is adapted and configured to: receive, from the first computing device the blinding function value and the set of data of the data source; transmit the blinding function value; receive, in response to the transmitting, an anonymous data identifier for the set of data; and store the set of data and the anonymous data identifier. The security node for the second computing device is configured and adapted to: receive, from the second computing device, the blinding function value for the set of data of the data source; apply a completion function to the blinding function value; generate the anonymous data identifier from the applying; and transmit the anonymous data identifier to the second computing device.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] For a fuller understanding of the nature and desired objects of the present invention, reference is made to the following detailed description taken in conjunction with the accompanying drawing figures wherein like reference characters denote corresponding parts throughout the several views.

[0016] FIG. **1** depicts an exemplary network diagram according to an embodiment of the invention.

[0017] FIGS. **2-5** depicts methods for secure sharing of data according to embodiments of the invention.

[0018] FIG. **6** depicts a system for secure data sharing according to an embodiment of the invention.

DEFINITIONS

[0019] The instant invention is most clearly understood with reference to the following definitions.

[0020] As used herein, the singular form "a," "an," and "the" include plural references unless the context clearly dictates otherwise.

[0021] As used in the specification and claims, the terms "comprises," "comprising," "containing," "having," and the like can have the meaning ascribed to them in U.S. patent law and can mean "includes," "including," and the like.

[0022] Unless specifically stated or obvious from context, the term "or," as used herein, is understood to be inclusive.

DETAILED DESCRIPTION OF THE INVENTION

[0023] Electronic patient medical records contain vast amounts of information of potential value to researchers striving to increase understanding of diseases, treatments, and outcomes. Effective use of such data is limited by privacy and technical concerns. Privacy laws require the removal of Personally Identifiable Information (PII) from the released data. Technical concerns are that the data must be abstracted for consistency across different providers. To be most useful, data from different providers for the same patient must be linked together. Embodiments of the invention apply cryptographic techniques to the problem of privacy-preserving linking of medical records.

INTRODUCTION

[0024] An embodiment of the invention facilitates a system for medical data sharing that meets several criteria: (1) understandable to stakeholders, (2) supportable by stakeholders, (3) simple and explainable, (4) actionable with minimal startup investment, (5) sustainable, and (6) secure.

[0025] Embodiments of the invention provide a new cryptographic primitive, called a blinding-completion pair, which addresses the practical problem of linking anonymous medical records. Blinding-completion pairs provide a method for generating a multitude of anonymous pseudonyms for an entity, to be used by data sources, and then consolidating that multitude into a single anonymous pseudonym at the destination database. Embodiments of the invention also provide a distributed system and protocol for sharing medical data that can preserve privacy when confronted with occasional data breaches.

Medical Information Workflow

[0026] Data enters the healthcare system when a patient contacts a provider, whether a primary care physician or a hospital. At that point, the provider determines and records the patient's PII and begins or updates the patient's chart.

[0027] To curate data for statistical and research purposes, trained registrars extract select data elements from the medical record according to specific data field definitions, resulting in highly structured data sets. The datasets are then stripped of PII and exported in a de-identified manner to one or more of several national databases. Importantly, not every healthcare entity submits to every database, and each database only requests certain fragments of the patient's medical information. As a result, each patient's care is captured by the databases in a piecemeal fashion. Because the databases do not collect PII, there is no way to consistently reunite the fragments of the healthcare data back together to create a complete picture of a patient's journey through the diagnosis, treatment, and outcome of their medical condition.

[0028] The analysis of patient outcomes captured within the databases has led to dramatic improvements in the safety and effectiveness of care for almost every medical condition. However, the ability of the database research to characterize relationships between variables and outcomes in medical care is critically dependent on the breadth of information available for analysis (e.g., to control for bias and confounding effects). Because databases are only capturing fragments of the medical journey, there are limitations to the types of improvements that currently can be made with database research.

[0029] For example, the database that best captures cancer stage does not capture the specific type of chemotherapy that patients received. If there were a way to reunite all the fragments of data back together, medical research using existing databases would become far more powerful, and many more improvements would be possible.

[0030] A simple but unacceptable "solution" to the linking problem is to give each patient a universal health identifier to be included with the patient's record in each database. This is used in some other countries (e.g., Norway). However, in the United States, the topic of a national identifier has become highly polarizing, making this a less feasible option. Moreover, anyone with access to the curated databases would be able to join the several databases into one master health record for the patient with that identifier. This is often sufficient, when combined with other information readily available on the internet, to deanonymize the health record and reveal the patient's PII.

Privacy-Preserving Linking of Patient Data

[0031] Embodiments of the invention provide a new cryptographic primitive for generating identifiers that allows a database to link records from different data providers while preserving privacy in the face of many kinds of breaches.

Blinding-Completion Pairs

[0032] Let x be the identifier used by a data provider h to identify an entity. Let b(x) be a one-way hash function. A blinding-completion pair for h is a pair of one-way functions ($b_h(x)$, $c_h(y)$) such that $b(x)=c_h(b_h(x))$ for all x, and b(x) is also one-way. Like a cryptosystem ($E_h(x)$,$D_h(y)$), the composition of the second function in the pair with the first yields the same function for all keys h. In our case, the composition is the fixed blinding function b(x), which defines the alias y=b(x) for x. While neither x nor y can be recovered from $y_h=b_h(x)$, y can be recovered from any single value y_h if the corresponding completion function c_h is available, since $y=c_h(y_h)$.

Implementation

[0033] There are several ways to implement blinding-completion pairs. One way is to use cryptographic accumulators. Let $Q=\{b_1, \ldots, b_N\}$ be a set of quasi-commutative cryptographic hash functions. They have the property that the N-way composition of these functions in any order yields the same function B. Hence, for any subset $S \subseteq Q$, the composition of those functions in S, call it b_S, can be used as the first element of a blinding-completion pair, and the composition of $b_{(Q-S)}$ becomes the completion function c_S. The drawback of this scheme is that N must be known in advance, and the time complexity of finding b_S and c_S grows with N.

[0034] Embodiments of the invention use a different scheme based on discrete logarithms. First, we introduce some standard number theory. For positive integer n, let Z^*_n be the set of positive integers less than n that are relatively prime to n. The size of Z^*_n is given by Euler's totient function $\phi(n)$.

[0035] In the special case that n is a prime p, $Z^*_p=\{1, \ldots, p-1\}$, so $\phi(p)=p-1$. Also, p has primitive roots. Say g is a primitive root of p if every number $a \in Z^*_p$ can be expressed as $a=g^k$ mod p for some $k \in Z^*_p$. The number k is called the discrete logarithm of a modulo p. Computing the discrete logarithm is believed to be computationally difficult when p and g are chosen carefully.

[0036] For our purposes, we choose $p=2q+1$, where q is a Sophie Germain prime and p is called a safe prime. Such prime pairs are widely used in cryptography, so a suitable supply exists for our purposes. An estimate of the number of Sophie Germain primes less than n is $\Theta(n/(\log n)^2)$.

[0037] There are $\phi(\phi(p))=\phi(p-1)$ primitive roots in Z. This makes it possible to find a primitive root g by a guess-and-check method. Guess a number $g \in Z^*_p$ and check that $g^q=-1 \pmod p$. The expected number of guesses required to find g is $(p-1)/\phi(p-1)=O(\log \log p)$. How big is $\phi(p-1)$? Because we've chosen $p-1=2q$, then $\phi(p-1)=\phi(2)\phi(q)=q-1=(p-3)/2$.

[0038] Let r, u be positive integers in $Z^*_{\phi(p)}=Z^*_{p-1}$, and let v be a positive integer less than $\phi(p)$ such that $r=(u+v)$ mod $\phi(p)$. Define $b_u(x)=xg^u$ mod p and let $c_v(y)=yg^v$ mod p. Then (b_u, c_v) is a blinding-completion pair for the blinding function $b(x)=xg^r$ mod p. This follows since

$$c_v(b_u(x)) \equiv c_v(xg^u) \equiv (xg^u)g^v \equiv xg^{u+v} \equiv xg^r \pmod p$$

The last identity follows from Euler's Theorem, which states that for a E Z^*_p, $a^{\Phi(p)} \equiv 1 \pmod p$.

[0039] The parameters p, q matter both for convenience and security. To choose r from $Z^*_{\phi(p)}$, we need to find an r that is relatively prime to $(p-1)$. But an arbitrary p−1 might have many small factors, e.g., p=71. However, since we choose p, q so that $p-1=2q$, we know the only factors of p−1 are 2 and q. Choosing a safe prime p makes it easy to find random numbers in $Z^*_{\phi(p)}$. As for security, the discrete logarithm problem is hard in general, but a solution may be feasible via the Pohlig-Hellman algorithm when p−1 has no large prime factors. A safe prime does not have this weakness.

[0040] As explained in the "Proposed Workflow for Enhanced Security" section below, one can allow security nodes to independently choose random values of r, $u \in Z^*_{\phi(p)}$ and calculate $v=(r-u)$ mod $\phi(p)$.

[0041] Because r and u may be chosen independently by different security nodes, there is a theoretical risk of v=0, which would produce the undesirable result $y_h=y$. From the point of view of a cryptographer, such a result is not a problem, but to satisfy our design requirements, we want to provide an unqualified guarantee that the identifier used by a data provider h does not appear in a database that links its records.

[0042] The value r is secret and may not be shared, therefore the security node choosing u cannot "peek" at r to make sure it chooses a safe value for u. In practice, when p is very large, the probability of r=u is vanishingly small. Should this ever happen, a simple remedy is to choose a new value for r via key rotation, as explained next.

Key Rotation

[0043] The values of r, u, v should be rotated periodically in case they are ever compromised. We provide a sketch of how such rotation could be implemented.

[0044] To rotate the value of r, choose a random $0<s<\phi(p)$ and calculate $r'=(r+s)$ mod $\phi(p)$. Check that $r' \in Z^*_{\phi(p)}$, and if not, choose a different random s and try again. Then recalculate $v'=(r'-u)$ mod $\phi(p)$. Finally, to update the blinded value $y=xg^r$ mod p, calculate a new blinded value $y'=yg^s$ mod p. Note that x is not needed for this calculation. Then y' is the new blinded value for the same x, since $y' \equiv xg^r g^s \equiv xg^{r'} \pmod p$.

[0045] To rotate the value of u, choose a random $0<s<\phi(p)$ and calculate $u'=(u+s)$ mod $\phi(p)$. Check that $u' \in Z^*_{\phi(p)}$, and if not, choose a different random s and try again. To update the blinded value $y_h=xg^u$ mod p, calculate a new blinded value $y'_h=y_h g^s$ mod p. Then y'_h is the new blinded value for the same x, since $y'_h \equiv xg^u g^s \equiv xg^{u'} \pmod p$.

Proposed Workflow for Enhanced Security

[0046] We propose three additions to the existing workflow to maintain security while still permitting research data sharing.

[0047] First, we envision a system of restricted local patient identifiers (LPIDs) that can be used to identify the medical records of a given patient within the context of a single healthcare provider. Local identifiers can be obtained from a patient's PII via a one-way cryptographic function. This prevents the local identifier from being reverse engineered to obtain PII.

[0048] Second, using the cryptographic technique of blinding-completion functions, the local patient identifiers for different health care providers can be used to calculate an anonymized patient identifier (APID). The APID allows a medical database to link patient records across providers while still providing no clear path to finding the corresponding PII.

[0049] Third, to further protect patient anonymity and privacy, we propose to separate the security services from the servers and databases holding the actual PII (in the case of hospitals) and medical data (in the case of curated database).

Trust

[0050] Our model of trust has two dimensions: whether the party has good intentions to keep sensitive information private, and whether the party is competent to do so. For example, while we may trust healthcare systems to do their best to keep sensitive patient information private, they are not always good at cybersecurity, as evidenced by the large number of cyber-attacks against health care organizations. Even when a healthcare system has a central information technology department capable of maintaining network security, that competency may be a scarce resource.

[0051] Our model of trust is different from traditional adversarial models that consider the worst possible outcomes from an untrusted party. Our model is informed by one author's experience in analyzing dozens of actual lawsuits related to online identity and privacy. While malice is sometimes present, incompetence is much more likely.

[0052] While healthcare providers are expected to be competent at medical treatment, there is no reason to expect them to be competent at cryptography (nor would we trust the average cryptographer to perform surgery). To mitigate the risk of healthcare providers performing cryptographic functions insecurely or leaking secret keys, embodiments of the invention can restrict access to certain functions and the secret keys that power them. To keep everyone safe, embodiments of the invention can introduce additional parties to the transaction, called "security nodes," which have demonstrated technical competence. Each healthcare provider can choose a security node to work with, and so can each medical database a depicted in FIG. **1**.

[0053] A security node is a network service that can be trusted to implement cryptographic functions correctly and to hold secret keys without leaking them. Security nodes could be independently operated, or they might be operated by a department within a medical organization with the required competence. Importantly, a security node can isolate the secret keys used by cryptographic functions or for signing messages in a single location. This makes it easier to protect secret keys by storing them in specialized computing hardware, such as a hardware security module (HSM). Security nodes also provide authentication services to health care providers. Each security node can have a public-private key pair it can use to sign and authenticate messages for other security nodes. A special "executive" security node can keep a list of all security nodes and their public keys. This list may be periodically updated and distributed, enabling security nodes to reliably authenticate each other's messages.

Parties

[0054] A transaction can include six parties: (1) a patient w, who is identified with a user identifier (UID), (2) a health care provider h, who treats patients and gathers medical data, (3) the health care provider's security node that provides LPIDs that can be attached to medical data in lieu of UIDs, (4) the database's security node that attaches an APID to medical data in lieu of LPIDs, (5) a database d that collects anonymous patient profiles, identified only by APID, and (6) researchers that receive anonymous patient profiles.

Identifiers

[0055] There can be three levels of identifiers, each with distinct properties:

[0056] First, UID can be an invariant identifier, such as a name-birthday pair or a Social Security number. The UID can be readily available to the patient and widely used by health care providers. A patient's UID is preferably not shared because it constitutes PII.

[0057] Second, LPID identifies patients relative to a healthcare provider and may have no apparent connection to any PII. A patient's LPID can be different at every provider and can be used for sending anonymized records to a database.

[0058] Third, APID identifies patients relative to a database and may have no apparent connection to any PII or to any LPID. Anonymized records sent to a database by different health care providers for the same patient can be associated with the same APID, which enables record linking.

[0059] The LPID and APID identifiers can be rotated periodically to frustrate any attacker who manages to breach the system. Rotation can be done if a breach is detected, or on a regular schedule to limit the damage from an undetected breach, and to provide other benefits.

Initialization

[0060] Initially, one or more databases join our proposed system, which provides the values p, q, and g. Each database d can choose a security node, which generates a random value r_d such that $r_d \in Z^*_{\phi(p)}$. The value r_d is used to generate blinding-completion function pairs and can be kept secret. Each healthcare provider h joining the system can choose a security node. The healthcare provider can obtain a public-private key pair for signing messages (e.g., an X.509 security certificate), using a digital signature algorithm such as DSA or ECDSA, and verifying the healthcare provider's identity to its security node. The public key can be registered, or "pinned," to the security node. Upon registration, the healthcare provider's security node can choose a random value u_h such that $u_h \in Z^*_{\phi(p)}$. The value u_h can be kept secret and is used to generate a blinding function $b_h(\ )$.

[0061] To join a database d, a provider h causes its security node to send u_h to the security node of database d. The security node of d then calculates a value $v_d = (r_d - u_h) \bmod \phi(p)$. The value v_d can be kept secret and is used to generate a completion function $c_h(\ )$. These blinding-completion functions can be constructed in such a way that: (1) each blinding function for each provider h produces a different pseudorandom identifier $LPID_h$ for each patient; and (2) each completion function for each provider h to each database d maps each $LPID_h$ to $APID_d$.

[0062] If a health care provider h participates in multiple databases, it uses the same $LPID_h$ identifiers, but each database d will generate different $APID_d$ identifiers. Conversely, when multiple providers contribute medical data to a database, each provider h has different $LPID_h$ identifiers and the database d has the same $APID_d$ identifiers. To preserve privacy, no provider knows any of the $APID_d$ values, and no database knows any of the $LPID_h$ values. The patient identifier equivalence pairs ($LPID_h$, $APID_d$) are only known to, or computed by, security nodes.

Contribution of Patient Profiles

[0063] Medical providers may contribute patient profiles to a database. A profile contains demographic and medical information of interest to researchers. For example, a patient profile might include age, medical diagnosis codes and dates, occupation, ethnicity, treatment history, and other target characteristics. Existing standards for storing digital medical records can be used.

[0064] To contribute a profile to a database d, a health care provider h can perform several steps.

[0065] First, the provider hashes the patient's UID, w, with a standard, widely available hash function such as SHA256, to generate a value x that it sends to its security node. The security node then applies the blinding function for that health care provider to x, resulting in the value $LPID_h$. The security node returns $LPID_h$ to the health care provider, and h adds it to the patient's medical record.

[0066] Second, the provider generates a random transaction number t. The relevant profile data m is then composed into a message (h, t, m, d) and sent to the database d. In some embodiments, the medical data will only be added to the database after it is authenticated by the database's security node.

[0067] Third, the provider creates a token (such as a JSON web token) containing the quadruple ($LPID_h$, h, t, d) and signs it using its secret key. The provider sends the signed token to the provider's security node, which then authenticates the signature using the health care provider's public key and appends its own signature to the token.

[0068] The provider's security node sends token ($LPID_h$, h, t, d) to the security node of database d which does the following steps: (1) authenticates the signature of the health care provider's security node; (2) verifies that the health care provider's name h in the token matches the name in the message; (3) applies the appropriate completion function c_h( ) to $LPID_h$ to generate $APID_d$, (4) creates a new token ($APID_d$, p, t, d), signs, and sends it to database d.

[0069] Database d receives the token ($APID_d$, h, t, d) and then performs these steps: (1) authenticates the signature of its own security node; (2) finds the message (h, t, m, d) with the same transaction number t; (3) verifies that the health care provider's name h in the message matches the health care provider's name in the token; and (4) adds the medical data m, the provider h, and $APID_d$ to its data store. If there is an existing record with $APID_d$, the new data is linked to the existing record.

Accessing Medical Data for Research

[0070] A researcher can connect to a database and search for patient profiles that match desired criteria for the study. Upon approval by an appropriate medical research ethics board, the researcher can then requisition specific medical data from the database that is relevant to the research being conducted.

[0071] Upon receiving an approved request for medical data related to a patient profile, the database then retrieves from its database the patient medical data that meets the researcher's specific criteria.

[0072] When releasing data to a researcher, the identifier for each record, $APID_d$, should be removed or hashed with a one way function such as SHA256. Researchers are not security experts. Therefore, they should not be trusted to keep the $APID_d$ identifiers private.

Threat Analysis

[0073] The system we describe, like all such systems, does not confer perfect security. If a security node were compromised, an attacker might learn the secret values r_d, u_h, or v_h. These secret values could enable an attacker to recover some or all of the blinding-completion functions b( ), b_h( ), or c_h( ) and their inverses. Having one or more of these inverse functions could give an attacker who possesses $APID_d$ the ability to calculate $LPID_h$ or x, the hash of the patient's UID. While it is not practical to invert the hash function used to generate x, an attacker could test whether a known UID value, when hashed, equals x.

[0074] Assume an attacker gathers medical data from researchers. This data would contain hashes of the $APID_d$ for each record. If the attacker additionally compromises health care providers and security nodes, it is conceivable that they could eventually link a UID to anonymous research data. Given UID and r_d, an attacker can calculate $APID_d$, hash this value and then compare it to the data collected from researchers. The difficulty of such an attack is high because it requires compromising at least one health care provider and at least one security node of a database containing data from that health care provider within a limited time frame (the key rotation period). Moreover, if such an attack were to succeed, it would likely deanonymize only a limited number of medical records, especially if there are many independent health care providers, databases, and security nodes.

[0075] The security nodes described in this system only need to communicate with other security nodes and with the medical providers or databases they serve. Consequently, a firewall can protect each security node so that it only communicates with systems on an "allow" list. This type of protection increases the difficulty of breaching a security node because even if the system has vulnerabilities, an attacker needs to gain access to a system on a security node's "allow" list even to commence a remote attack on the security node.

[0076] We believe that the difficulty of attacking our proposed system is sufficiently high, and the profitability sufficiently low, that attackers would prefer to attack health care providers directly and aggregate data via UID. Therefore, our proposed system does not materially increase the risk of private medical data being exposed in a data breach versus the status quo.

CONCLUSION

[0077] We have presented a new cryptographic technique called blinding-completion pairs and demonstrated how they could be used to enable the sharing of private data without revealing personally identifiable information (PII).

[0078] Based upon blinding-completion pairs maintained by security nodes, we have drawn a sketch of how health care providers could supply medical data to one or more databases that would aggregate data for each patient and then make those consolidated records available as anonymous data to researchers. Our system could release data for medical research in a way that protects patient PII while still enabling qualified researchers to identify records from different health care providers that belong to the same patient.

[0079] Possible areas for future work include constructing a prototype system, developing new blinding-completion functions with improved security properties, and investigating alternative sharing protocols that may offer stronger privacy guarantees in the event of data breaches.

EQUIVALENTS

[0080] Although preferred embodiments of the invention have been described using specific terms, such description is for illustrative purposes only, and it is to be understood that changes and variations may be made without departing from the spirit or scope of the following claims.

INCORPORATION BY REFERENCE

[0081] The entire contents of all patents, published patent applications, and other references cited herein are hereby expressly incorporated herein in their entireties by reference.

1. A method for secure sharing of data, the method comprising:

receiving, from a first computing device and by a security node for the first computing device, a hashed identifier for a data source;

generating, in response to the receiving, a blinding function value dependent on the hashed identifier; and

transmitting, to the first computing device, the blinding function value for storage of a set of data and linking the set of data to the data source.

2. The method of claim **1**, further comprising:

applying a blinding function to the hashed identifier to generate the blinding function value.

3. The method of claim **2**, wherein the blinding function is of the form $b_u(x)=xg^u$ mod p, wherein:

x comprises the identifier for the data source;

g comprises a primitive root value;

u comprises a secret random value selected by the security node for the first computing device; and

p comprises a safe prime value.

4. The method of claim **1**, wherein the first computing device comprises a device for a healthcare provider.

5. The method of claim **1**, wherein the data source comprises a medical patient.

6. A method for secure sharing of data, the method comprising:

transmitting, from a first computing device and to a security node for the first computing device, a hashed identifier for a data source;

receiving, in response to the transmitting, a blinding function value dependent on the hashed identifier; and

transmitting, from the first computing device to a second computing device, the blinding function value and a set of data of the data source for storage of the set of data and linking the set of data to the data source.

7. The method of claim **6**, further comprising:

generating the hashed identifier by applying a hashing function to an identifier of the data source.

8. The method of claim **7**, wherein the hashing function comprises a SHA256 hashing function.

9. The method of claim **7**, wherein the hashing function comprises a one-way hashing function.

10. The method of claim **6**, wherein the second computing device comprises a database configured for storing medical records.

11. A method for secure sharing of data, the method comprising:

receiving, from a second computing device and at a security node for the second computing device, a blinding function value for a set of data of a data source;

applying, by the security node, a completion function to the blinding function value;

generating an anonymous data identifier from the applying; and

transmitting the anonymous data identifier to the second computing device for storage of the set of data and linking the set of data to the data source.

12. The method of claim **11**, wherein the completion function is of the form $c_v(y)=yg^v$ mod p, wherein:

y comprises a hashed identifier for the data source;

g comprises a primitive root value;

v comprises a secret random value selected by the security node for the second computing device; and

p comprises a safe prime value.

13. A method for secure sharing of data, the method comprising:

receiving, from a first computing device and at a second computing device, a blinding function value and a set of data of a data source;

transmitting, form the second computing device to a security node of the second computing device, the blinding function value;

receiving, in response to the transmitting, an anonymous data identifier for the set of data; and

storing the set of data and the anonymous data identifier at the second computing device.

14. The method of claim **13**, further comprising:

identifying another set of data stored at the second computing device with the anonymous data identifier; and

linking the set of data to the other set of data based on the identifying.

15. A system for secure data sharing, the system comprising:

(a) a first computing device configured and adapted to:

transmit a hashed identifier for a data source;

receive, in response to the transmitting, a blinding function value dependent on the hashed identifier; and

transmit the blinding function value and a set of data of the data source for storage of the set of data and linking the set of data to the data source;

(b) a security node for the first computing device and configured and adapted to:

receive, from a first computing device, the hashed identifier for the data source;

generate, in response to the transmitting, the blinding function value dependent on the hashed identifier; and

transmit, to the first computing device, the blinding function value;

(c) a second computing device adapted and configured to:

receive, from the first computing device the blinding function value and the set of data of the data source;

transmit the blinding function value;

receive, in response to the transmitting, an anonymous data identifier for the set of data; and

store the set of data and the anonymous data identifier; and

(d) a security node for the second computing device configured and adapted to:

receive, from the second computing device, the blinding function value for the set of data of the data source;

apply a completion function to the blinding function value;

generate the anonymous data identifier from the applying; and

transmit the anonymous data identifier to the second computing device.

* * * * *