Authentication With Block-Chain Algorithm and Text Encryption Protocol in Calculation of Social Network

Seminar Report

submitted in partial fulfillment of the requirement
for award of Degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

of

APJ Abdul Kalam Technological University

Submitted by

ANUSREE P S

Department of Computer Science and Engineering
Mar Athanasius College of Engineering
Kothamangalam

Authentication With Block-Chain Algorithm and Text Encryption Protocol in Calculation of Social Network

Seminar Report

submitted in partial fulfillment of the requirement
for award of Degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

of

APJ Abdul Kalam Technological University

Submitted by

ANUSREE P S

Department of Computer Science and Engineering
Mar Athanasius College of Engineering
Kothamangalam

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MAR ATHANASIUS COLLEGE OF ENGINEERING
KOTHAMANGALAM

CERTIFICATE

This is to certify that the report entitled **Authentication With Block-Chain Algorithm and Text Encryption Protocol in Calculation of Social Network** *submitted by* **Ms. ANUSREE P S, Reg. No.MAC15CS014** *towards partial fulfillment of the requirement for the award of Degree of Bachelor of Technology in Computer science and Engineering from APJ Abdul Kalam Technological University for December 2018 is a bonafide record of the seminar carried out by her under our supervision and guidance.*

....	....	....
Prof. Joby George	**Prof. Neethu Subash**	**Dr. Surekha Mariam Varghese**
Faculty Guide	***Faculty Guide***	***Head Of Department***

Date:

Dept. Seal

ACKNOWLEDGEMENT

First and foremost, I sincerely thank the 'God Almighty' for his grace for the successful and timely completion of the seminar.

I express my sincere gratitude and thanks to Dr. Solly George, Principal and Dr. Surekha Mariam Varghese, Head Of the Department for providing the necessary facilities and their encouragement and support.

I owe special thanks to the staff-in-charge Prof Joby George, Prof Neethu Subash and Prof Joby Anu Mathew for their corrections, suggestions and sincere efforts to co-ordinate the seminar under a tight schedule.

I express my sincere thanks to staff members in the Department of Computer Science and Engineering who have taken sincere efforts in helping me to conduct this seminar.

Finally, I would like to acknowledge the heartfelt efforts, comments, criticisms, co-operation and tremendous support given to me by my dear friends during the preparation of the seminar and also during the presentation without whose support this work would have been all the more difficult to accomplish.

ABSTRACT

Community detection is an important aspect of social network analysis. Most of the existing methods are single classification algorithms. Multi-classification algorithms that can discover overlapping communities are still incomplete. The aim is to propose an efficient algorithm to preserve the privacy of information in social networks. During expansion of communities user's identities are verified after they send request. Block chain is used to store the user's public key and bind to the block address. In order to prevent the curious users from illegal access to other user's information, plain text is not sent directly after authentication. Instead the attributes are hashed by mixed hash encryption to make sure that users can only calculate matching degree. This method thereby avoids requirement of third party authentication and add more privacy to user data.

CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATION

CMCR	Community Mining and Content Recommendation
CSE	Closeness Seed Expansion
ISCE	Influence Social Community Expansion
KSM	K-clique-community Seed Mining
MCRA	Message Content Recommendation Algorithm
PoW	Proof of work
RSA	Rivest–Shamir–Adleman
LDA	Latent Dirichlet Allocation

Introduction

With the rapid development of mobile Internet, large-scale mobile social network has become the most popular platform for communication and information propagation. Privacy in social networks is a large and growing concern in recent times. It refers to various issues in a social network which include privacy of users, links, and their attributes. Each privacy component of a social network is vast and consists of various sub-problems. For example, user privacy includes multiple sub-problems like user location privacy, and user personal information privacy.

Online social networks have gained huge popularity in today's world. People tend to use multiple networking sites such as Facebook, Twitter, and LinkedIn. Social networks have become an inevitable part in our daily life. We store information such as date of birth and exact place of ones' location in social media. We store all our personal content online such as contacts, photos, and bookmarks. We also interact with many people through posts, tweets, and tags. Various types of social networks are used in present day, and each one has a unique feature.
1) Personal networks: These type of networks focuses on creating the detailed online profile of an individual by including as many user attributes as possible. They also allow users to connect. Websites like Facebook, Friendster, and MySpace are some examples.
2) Status update networks: These type of networks focuses on posting user updates online. These posts include user's interests, places visited and their personal thoughts. Twitter is the best example of this type.
3) Shared-interest networks: These type of networks is designed to get a group of people together with common interests. For example, LinkedIn is a professional network website developed for employers and job seekers. Research-Gate is developed for researchers to broadcast their publications and contact the authors of their papers of interest.
4) Neighborhood Exploring networks: These type of networks is designed to find users in the neighborhood for sharing information, media files, and interaction. These communications may later lead to a personal meeting and hence searching for neighbors depend hugely on location information.

Recent development in the technology has made it easier to collect massive amounts of social network data and hence leads to serious privacy concerns. For example, in networks like LinkedIn, an intruder who necessarily is not a friend of a person, can also gather information about how he is related to that person. He can collect information on common friends or a link of friends between him and the person he intends to attack.

1) Behavioral advertising is an example of privacy breach. These advertisements are tailored based on person's interest. These kinds of ads are much more profitable in social networks as the probability of a user clicking on these ads are higher than untailored ads. Google, Facebook, Twitter and all major social networking sites provide access to the third

party advertising companies. These advertising companies use a variety of data such as the location of the user, relationship status and so on. Even on networks like Facebook, we see advertisements of travel agencies if we change the status to married. Data released is anonymized but sometimes unintentionally leaks some of the user's crucial information. Facebook has exposed six million users' phone numbers and email addresses to unauthorized users for a year.
2) Identity theft uses an individual's personal information often for financial gain. The information posted on users' social network profiles is used to steal their identity. In 2009, researchers at Carnegie Mellon University had found that the data extracted from social networks and other online public databases can lead to the discovery of partial or full social security numbers.
3) Stalking or Child Abuse One of the early privacy cases in this regard occurred in 2010 on MySpace where minors were bullied and led to the adoption of "age requirements and other safety measures". Events such as stalking and "catfishing" are frequent in present society and hence has become a prominent topic in social network security.

Social network is a network of social interactions and personal relationships. Any social network can be represented as a graph, where each user's profile is a node and friendship between two users is an edge between those two nodes(Fig 1.1).Users connect with similar users forming communities.Certain users belong to multiple communities forming overlapping communities.

Fig. 1.1: Social network with various users and shared attributes.

It is observed that in complex social networks, there are multiple overlapping communities. Numbers of community detection methods exist, one of which is to detect communities by analyzing user properties and interests. However, as social networks have become more and more complex, we can no longer detect communities according to these simple rules. To deal with that, in user centric social networks, we detect social communities according to user influence, user relation and interaction, also we practice personalize recommending based on semantic analysis and statistical analysis. Finding and analyzing community structure often provide invaluable help in deeply understanding the structure and function of a network.

Social network applications reflect real world to cyberspace, thus lead to privacy leaks while detecting social circles according to user information. But once this privacy information is illegally obtained, criminals will be able to obtain the relationship between users, infer other user interests, and use other user personal information. To avoid that risk, we need to validate the identity of user in the process of building social circles by relationship of users.

In this process, public key cryptography is usually used. But this method would fail on one condition: when someone forges the identity and key to match, the user cannot be identified effectively, resulting in his private information being leaked. Digital certificates can solve this problem, but publishing digital certificates require specific agencies, which make them inconvenient to use. Especially in the social network, since the amount of users is huge, it is impractical to make a digital certificate for each person.

Therefore, the use of block chaining technology to protect the identity of users and the security of their corresponding keys, so as to resist the attacks from semi-honest users and malicious users is proposed. CMCR algorithm is improved through two aspects i.e. user authentication and text encryption, so that it can better ensure the security of user information. In terms of user authentication, the Authentication with Block-chain algorithm to verify the identity of user is proposed. In terms of recommendation text, they proposed a protocol combining RSA algorithm to prevent users illegally acquiring information.

Related Works

In dealing with social communities detection, most of early approaches are based on the entire network structure, so they don't function well on detecting overlapping communities of a certain user. And most of these algorithms are classification algorithms. The methods of community detection, and the features of overlapping communities are represented in Fig 2.1. Although he focus is on the overall structure of the network, it does not provide a way to detect the community of a particular user. An algorithm named CMCR for a user centric network, based on Clique Theory, PageRank, and LDA, is used to solve the problem. Experimental results show that the proposed algorithm can outperform baseline algorithms in some common criteria. However, when designing this algorithm, there is a slight lack of user privacy and user information protection.

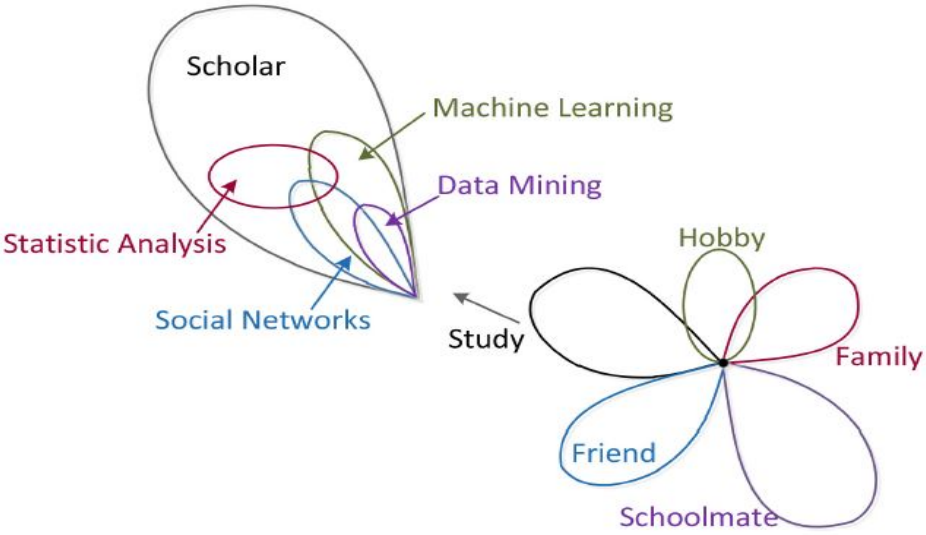


Fig. 2.1: The overlapping communities

There are two main privacy protection mechanisms in user profile based user matching. A mechanism treats user profiles as collections of attributes and then matches them according to the set of attributes. Another mechanism allows user profiles to be matched by vector dot products by regarding as user profiles as vectors. The two mechanisms mentioned above are based on public key cryptography and homomorphic encryption technology, so the computation cost is very high. And these two mechanisms require a trusted third party organization, which is hard to implement in social networks. We do not use preconfigured trusted third party organizations to encrypt using attribute information, which is not a direct match between users.

2.1 Clique and community

Cliques are formed with complete graphs. Two k-cliques are adjacent if they share k-1 nodes(Fig 2.2.)

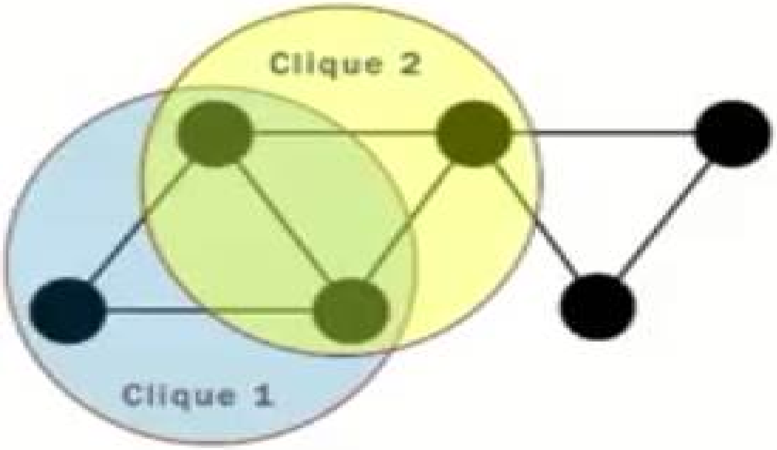


Fig. 2.2: clique with k=3

A community is defied as the maximal union of k-cliques that can be reached from each other through a series of adjacent k-cliques(Fig 2.3). Detecting community involves finding densely connected nodes.

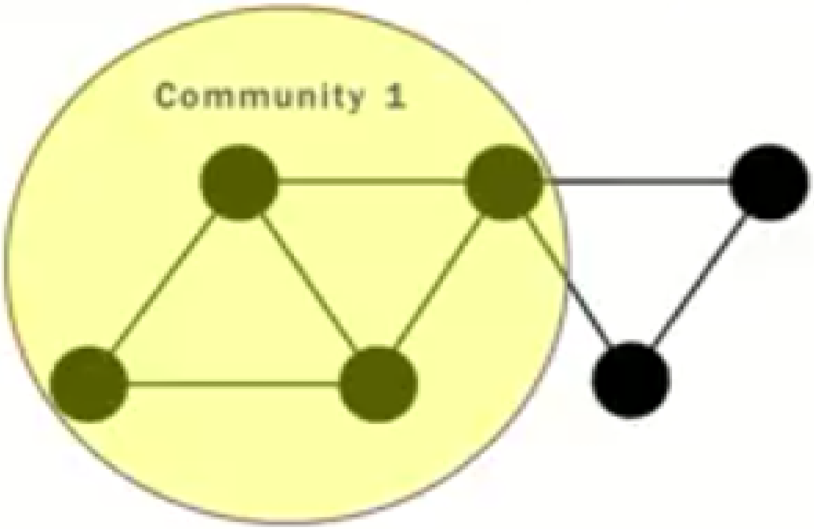


Fig. 2.3: community

It reveal the structure of social network, dig to people's view, analyze the information dissemination, and grasp as well as control public sentiment. A good community should be internally well-connected and also well-separated from the rest of the network. Communities in networks also overlap as nodes belong to multiple clusters at once(Fig 2.4)

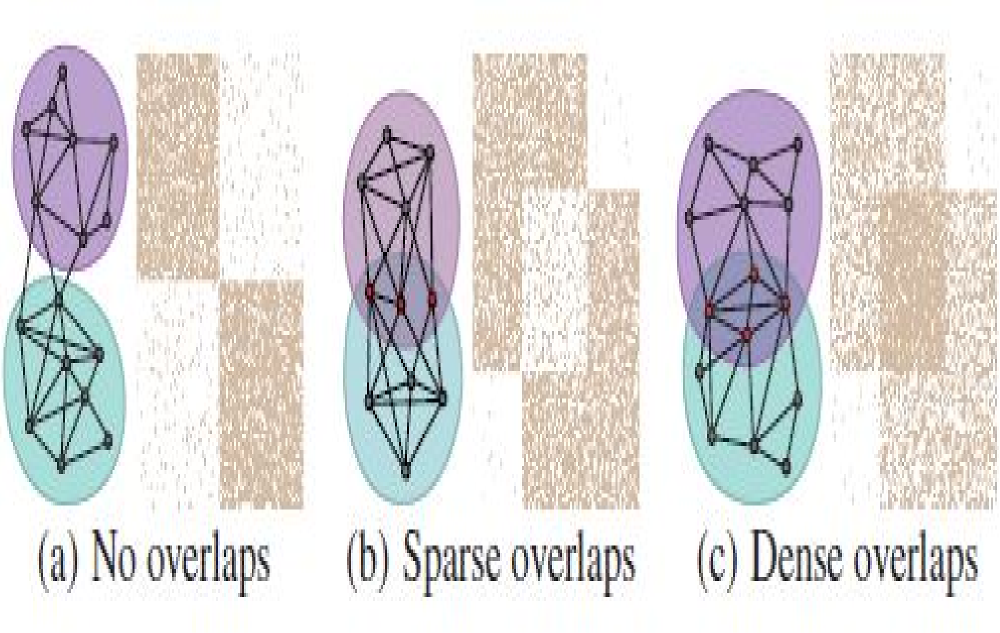


Fig. 2.4: Three views on the structure of network communities. Left: Network; Right: Corresponding adjacency matrix.

Due to the difficulties in evaluating the detected communities and the lack of scalable algorithms, the task of overlapping community detection in large networks largely remains an open problem. the more communities a pair of nodes shares the more likely they are connected in the network. For example, people sharing multiple hobbies (i.e., interest based communities) have a higher chance of becoming friends, researchers with many common interests (i.e., many common scientific communities) are more likely to work and publish together.

2.2 Privacy types in social networks

There are three categories in which the privacy has to be maintained.

Node Privacy: Information about the user has to be preserved.

Attribute privacy: Information about user attributes like age, sex, interests, location and so on should be preserved.

Link privacy: Information about the link or friendship between users should be preserved.

2.2.1 Node privacy

In this kind of anonymization, we aim to anonymize the user in a network which is same as the anonymization of the vertices in a graph. This anonymization can be achieved by a simple technique called Naive anonymization. In this method, all the vertices are replaced with random numbers or alphabets. Thus, an attacker retrieving a node perceives a number but does not have the referencing physical node information.This kind of anonymization is effortless, but is prone to simple attacks like infiltration, where the intruder extracts the information by adding himself as a friend of certain nodes and retrieve information on all the links that go through these nodes. For example, in networks like Facebook, it is easy to retrieve information of “Friends of Friends” if the profiles are made public. This type of crawling may lead to the discovery of complete or partial network and thus leading to the development of the network structure locally.

An attack, called the walk-based attack, had been proposed to understand the connection between any two given nodes. In this method, an attacker creates k different accounts and links them randomly in the network. He then creates a specific link pattern with the nodes of interest. Once the adversary establishes these connections, it is easy to identify the subgraph of the nodes, in the anonymized graph, that corresponds to his accounts with a high probability.we measure the following quantities for each edge:

Closeness Centrality: This is the average shortest path from one node to all the nodes in the graph.

Betweenness Centrality: This is the proportion of all shortest paths through a given node.

Path Length Distribution: This is constructed from all the shortest paths between each pair of nodes.

Degree Distribution: This is the distribution of all the vertices degrees

Diameter: This is the maximum shortest path between any two given nodes. Perturbation is one of the common techniques used to add or delete node into the graph. But perturbation, if exceeded more than a certain level, can cause randomness and leads to information loss.

2.2.2 Link privacy

A link in a social network can be presumed as a connection from one user to the other. One of the serious concerns in social network privacy, targeted advertisements, uses sensitive attributes like ones’ health information, bank name, country of origin and so on for their data mining. A recent article shows that Facebook provides user categories to third party advertisers. This leak of information may cause the loss of private information. Link privacy can be achieved with a simple technique called “Edge Perturbation”. Primitive edge anonymization can be achieved in five different ways:

Intact edges: This technique removes sensitive edges leaving all the other edges intact.

Partial-edge removal: This technique removes certain percentage of observations based on a prespecified criterion. For example, removing edges that connects high degree nodes at random.

Cluster-edge anonymization: In this method, we form clusters of different edge types and make sure the number of edges between these clusters remain same, even after removing and adding random edges.

Cluster-edge anonymization with constraints: This technique is an extension to the previous technique. Here, we assume additional constraints like any two equivalence classes should have same limitations to the corresponding nodes in the original graph.

Removed edges: This technique removes all edges.

2.3 Block chain

Metaphorically, it can be described as a chain of blocks that contain information.It was invented in 1991 to time-stamp digital documents like a notary. Later it was adapted by Satoshi Nakamoto in 2009 to create digital cryptocurrency bit coin. Block chain is a decentralized & distributed ledger that is completely open to anyone. To change data in a particular block you don't rewrite it, instead the change is stored in a new block. It creates trust in the data through proof-of-work. Before adding a block it's proof-of-work needs to be calculated and shared throughout the network.

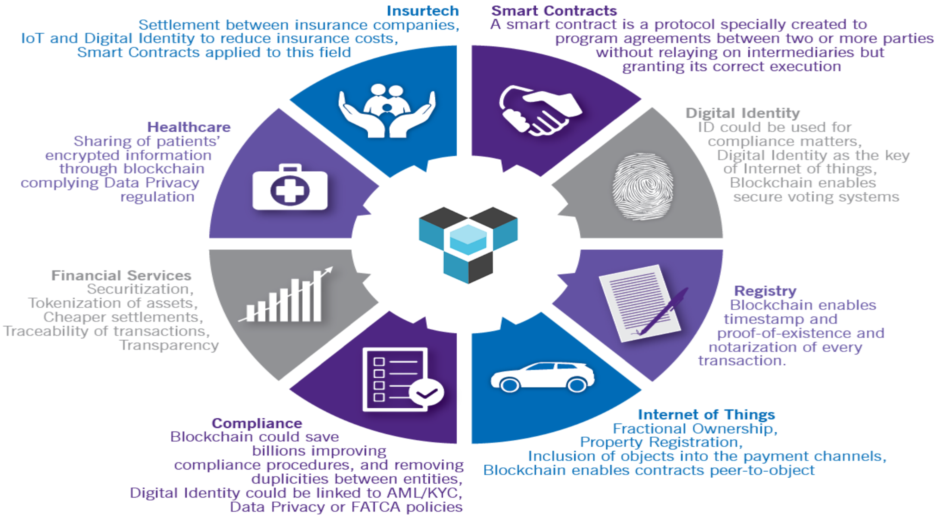

Insurtech
Settlement between insurance companies,
IoT and Digital Identity to reduce insurance costs,
Smart Contracts applied to this field
Smart Contracts
A smart contract is a protocol specially created to
program agreements between two or more parties
without relaying on intermediaries but
granting its correct execution
Healthcare
Sharing of patients'
encrypted information
through blockchain
complying Data Privacy
regulation
Digital Identity
ID could be used for
compliance matters,
Digital Identity as the key
of Internet of things,
Blockchain enables
secure voting systems
Financial Services
Securitization,
Tokenization of assets,
Cheaper settlements,
Traceability of transactions,
Transparency
Registry
Blockchain enables
timestamp and
proof-of-existence and
notarization of every
transaction.
Compliance
Blockchain could save
billions improving
compliance procedures, and removing
duplicities between entities,
Digital Identity could be linked to AML/KYC,
Data Privacy or FATCA policies
Internet of Things
Fractional Ownership,
Property Registration,
Inclusion of objects into the payment channels,
Blockchain enables contracts peer-to-object

Fig. 2.5: Block chain

Block chain is a kind of technique realization of the electronic currency book system by peer-to-peer, it can record every bitcoin transaction records without a center server in a network system, and it is maintained by participants. No one can change the contents of the block chain without authorization, thus it has very stable security for its holder. It allows any two users to trade directly without a trusted third party mechanism. The block chain records all transactions that occur in the bitcoin system, and once the transaction information is recorded, it is permanently stored and cannot be changed.Since most nodes are controlled by honest network nodes, attacks are very difficult to implement, thus the block information in the block chain is trustworthy. The anonymity of participants in the bitcoin system ensures the security of their privacy. Participants can either voluntarily leave or re-enter the bitcoin system by receiving the longest workload proof chain to obtain transaction information that occurs when leaving the system Fig(2.5).

Block chain technology and distributed Sub Ledger have attracted much attention and led to many projects in different industries[6]. Paper [7] uses block chaining technology to run the cash system in a peer to peer environment and prevents double payment, and solves the problem that a trusted third party handle the electronic payment information. Paper [8] applies block chain technology to social networks, uses reliability scoring to improve the system, and analyzes attacks rather than using PoW(proof of work) to protect them. But in the process of encrypting configuration file, RSA is the first public key cryptosystem to prevent applications, and its security has always been the focus of cryptography research. It is widely used in various applications in the security field. RSA and other related technologies, combined with biological development of new technologies have become a new research point[9].

2.4 Attack models

The two main adversary models that have been considered are semi-honest adversaries who follow the prescribed protocol but try to glean more information than allowed from the protocol transcript, and malicious adversaries who can run any efficient strategy in order to carry out their attack.One would naturally expect that any protocol that is secure in the presence of malicious adversaries will automatically be secure in the presence of semi-honest adversaries. However, due to a technicality in the definition, this is not necessarily true.

1) SEMI-HONEST MODEL In this model, semi honest members are also called passive attackers. In the process of multi-party computation, a semihonest member fully abides by the implementation of the agreement, neither withdraws from the agreement, nor tampers with the results of the protocol. He or she may retain some intermediate results in the implementation of the agreement and attempt to analyze and derive input data from other members through these intermediate results.

2) MALICIOUS MODEL In the model, malicious attackers are also active attackers. In the calculation process, a malicious attacker cannot follow the protocol process execution, interrupt protocol operation process, and collude with the intermediate results or modify the agreement with other parties.

2.5 Social community mining and content recommendation

First Level Relation Graph G1:
G1 = {V1,E1}, V1 = {n|c follows n}, E1 = {(n1, n2) |n1 follows n2, n1 ∈ V1, n2 ∈ V1}.

Second Level Relation Graph G2:
G2 = {V2,E2}, V2 = {n|user followed by users in V1} ∪ V1, E2 = { (n1, n2)|n1 follows n2, n1 ∈ V2, n2 ∈ V2}.

Collective Friends: Define CF(n1, n2) as a set of common friends between User n1 and User n2.

Closeness: Define CL(n1, n2) as the number of Collective Friends between User n1 and User n2. CL(n1, n2) = |CF(n1, n2)|.

Closeness Distance: Define D(community, n) as the closeness distance from a certain User n to a certain social Community community, as shown in the Equ 2.1, i ∈ community, s = |community|.

$$D(community, n) = \frac{\sum |CF(n,i)^{CF(n,c)}|}{s * |CF(n,c)|} \qquad \text{(Equ:2.1)}$$

In the definitions above, V is the set of vertices, which represents the set of users. E is the set of edges, which represents the relations. c stands for the centric user and n is an user in microblog.

The Proposed Method

3.1 Perfection of social circle detection

Overlapping community detection would be done by mining seeds and expansion. Social circle expansion would be conducted by extending the seed set Seed(c) through two algorithms of Closeness Seed Expansion (CSE) and Influence Social Community Expansion (ISCE) algorithm. According to the user information and the user relationship, seed mining can obtain the seed set Seed(c) by an algorithm named K-clique-community Seed Mining (KSM) algorithm to solve that problem.

3.1.1 Mining seed sets using KSM

The problem of mining seed sets means to locate some potential foundations for social communities given the users and relations.KSM takes the two important theories of Clique Theory into account to implement the mining. The theories are summarized as follows:

- For any clique with the size s, s $>$ k, it forms a community itself.
- For any two cliques with a overlapping part whose size $\geq$ k-2, they form a community together.

k means the minimum threshold of the size of a community to be mined. Given the users and relations, KSM first locates all the max cliques. With all the cliques detected, the Clique-Overlapping Matrix M is built where M_{ij} stands for the number of public nodes shared by Clique i and j. M_{ii} is the size of Clique i. Then based on the two theories above, the adjacency matrix M' of the undirected graph is computed. The Seed(c) would be the result of a depth-first search of connected sub-graph in M'.The pseudocode of KSM algorithm is shown in Algorithm 1.

3.1.2 Overlapping Communities Detection

The overlapping communities of centric user are detected by expanding Seed(c).

Algorithm 1 K-clique-community Seed Mining

```
    ▷ %input: centric user c, the First Level Relation Graph of centric user G1,parameter k%
        ▷ %output: the set of seeds of social interest circle of specific user Seeds(c)%
Seeds(c) ← $\phi$
search all the max cliques in $G_1$;
build Clique-Overlapping Matrix M according to the result of step 2;
for M, subtract the diagonal elements by k−1, the else k−2;
for each element e of M do
    if (e<0) then e←0;
    end if
    if (e>1) then e←1;
    end if
end for
M’ ← M
Seeds(c) ← depth-first search connected sub-graph on M;
return Seeds(c);
```

Closeness Seed Expansion (CSE)

It adopts a greedy strategy of considering user with higher closeness to centric user prior. CSE algorithm uses seeds as initial social communities, and makes greedy expansion based on the feature of closeness.The pseudocode of CSE algorithm is shown as Algorithm 2. Nodes closer enough to a community would be added to it first. Parameter k in K-clique-community is used as standard to filter out the social circle with size not larger than k.

Influence Social Community Expansion (ISCE)

The second algorithm used in social circle expansion is ISCE algorithm, which focuses on the user’s impact on the network. First, we calculate the user’s impact, according to which the candidates are sorted in the second-level relationship in the ISCE. The Calculation process of user impact is shown in the Fig 3.1. Then the classic modular function Q is adopted as the standard. If the present social structure would form a better social structure with a new added node, the modularity will increase. Then that node would suit the community for sure. By analysis, we believe that the number of users replying to a message can indicate his interest in message.The pseudocode of ISCE algorithm is shown as Algorithm 3.

Algorithm 2 Closeness Seeds Expansion

```
                                        ▷ %input: c, $G_1$, Seeds(c), k%
                                        ▷ %output: Community($G_1$)%
Community($G_1$) $\leftarrow$ Seeds($G_1$);
Candidates($G_1$) $\leftarrow$ find candidate nodes not included in seeds in $G_1$;
for every node n in Candidates($G_1$) do calculate Closeness(n, c);
end for
sort nodes in Candidates($G_1$) by descending order of closeness;
for every node n in Candidates($G_1$) do
    for every circle c in Circles($G_1$) do d $\leftarrow$ calculate D(c, n);
        if d $\geq \delta$ then add node n into circle;
        end if
    end for
    if n has not been added to any circles then Community($G_1$) $\leftarrow$ build a new community with n;
    end if
end for
for every community c in Community($G_1$) do
    if c.size()$\leq$ k then delete c from Community($G_1$);
    end if
end for
```

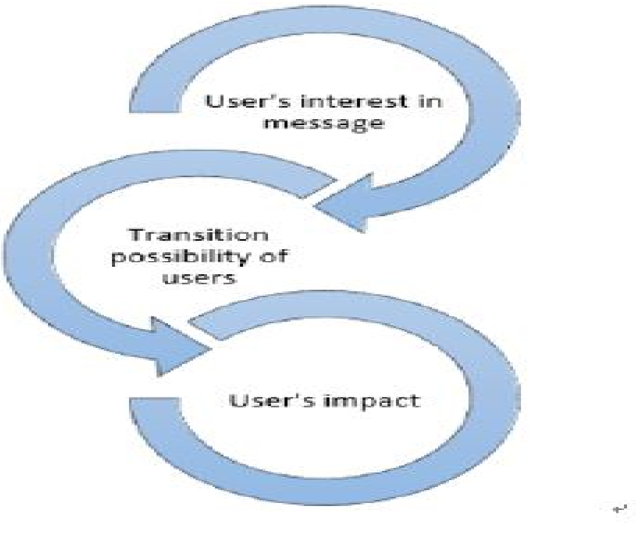


Fig. 3.1: Calculation process of user impact..

To calculate the influence of users, a similar idea of PageRank is utilized to represent the following relations of users. The key point of PageRank is the transition possibility of nodes in Random Walk Model.

Algorithm 3 Influence Social Circle Expansion

```
                                              ▷ %input: $G_2$, Community($G_1$)%
                                                ▷ %output: Community($G_2$)%
Community($G_2$) ← Community($G_1$);
Q ←calculate the modularity of $G_2$ with Community($G_2$);
calculate influence PR(i) for every node of $V_2$ by extended PageRank algorithm;
Candidates($G_2$)← find nodes not in any social interest circles in $V_2$;
sort nodes in Candidates($G_2$) by descending order of influence;
for each node in Candidates($G_2$) do
    for each social interest circle in Community($G_2$) do add node into circle; Q' ← calculate the modularity of the community after adding that node;
        if Q'< Q then delete node from community;
        end if
        if Q'> Q then Q ← Q';
        end if
    end for
end for
return Community($G_2$);
```

We First, the similarity between users Sim(i, j) is calculated by cosine similarity of their Message Interest Feature Vector Equ(3.1), where id stands for the ID of a message, and cn(u, id) means the number of replies that User u replies to Message id, under the assumption that the number a user replies to a message can represent his interest degree to that message.

Message Interest Feature Vector: Define V (u) as message interest feature vector:

$$V(u) = [id_1 = cn(u, id_1), id_2 = cn(u, id_2), ..., id_t = cn(u, id_t)], t = |M(u)|, id_i \in M(u) \tag{Equ:3.1}$$

Given the similarities of users, the redefined calculation method of transition possibility of users is shown as Equ 3.2. It actually stands for the ratio of information which User i is interested in, from the whole information he gains. The numerator is the quantity of information User i receives from User j. The denominator is the whole quantity of information User i receives.

$$P_{ij} = \frac{M(j) * Sim(i, j)}{\sum_{n \in V_1} |M(n)||Sim(i, n)|}, i, j \in V_2 \tag{Equ:3.2}$$

With the redefinition of transition possibility above, the influence degree of a user could be calculated by Equ. 3.3 where q is set by experience of PageRank.

$$PR_i = \frac{1-q}{V_2} + q \sum_j PR(j) * P_{ij}, i, j \in V2, q = 0.85 \quad i, j \in V_2 \tag{Equ:3.3}$$

3.2 Authentication and relationship encryption

CSE algorithm is based on the intimacy of users. In sociology, if two people have more common friends, the relationship between the two users is more intimate. So we calculate intimacy by calculating two users who follow each other. However, in social network implementation environment,to conduct social circle detection, a user would have to know other users' private information, in order to prevent other people illegally obtain user information, any two party users should be verified their identity before the process of communication. Therefore, before a user acquires information of the other user relationship, we need to authenticate identity of that user, in case a malicious user gets the relationship of the users illegally, and infer the user interests and preferences.

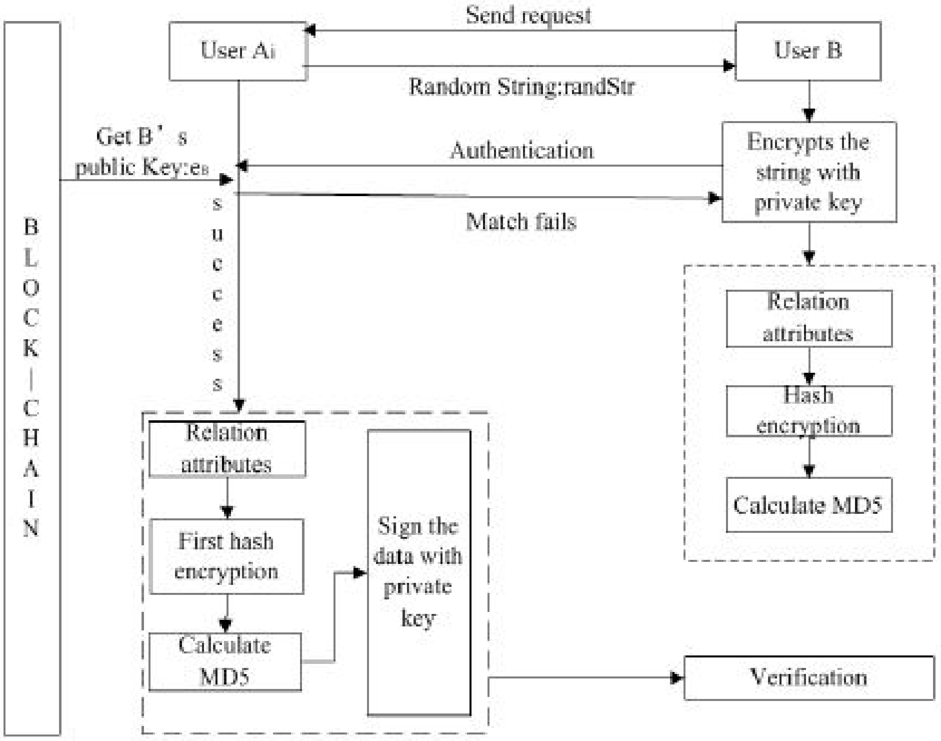


Fig. 3.2: Authentication model flow.

The Authentication with Block-chain algorithm is used to authenticate users. A pair of private key and public key is created for each user. To ensure the security of the key, the public key chain is stored according to the block chain constructed by the bitcoin system. The model flow is shown in Fig 3.2.

3.2.1 RSA algorithm

RSA is one of the first public-key cryptosystems and is widely used for secure data transmission.RSA was first described in 1977 by Ron Rivest, Adi Shamir and Leonard Adleman of the Massachusetts Institute of Technology. In such a cryptosystem, the encryption key is public and it is different from the decryption key which is kept secret. The algorithm is based on the fact that finding the factors of a large composite number is difficult: when the integers are prime numbers, the problem is called prime factorization. It is also a key pair (public and private key) generator. It provides a method of assuring the confidentiality, integrity, authenticity and non-reputability of electronic communications and data storage. The public and the private key-generation algorithm is the most complex part of RSA cryptography.

In generation of public key and private key, we use the state-of-art RSA algorithm.

Two large prime numbers, p and q, are generated using the Rabin-Miller primality test algorithm. Suppose N=p^rq. This number is used by both the public and private keys and provides the link between them. Its length, usually expressed in bits, is called the key length. The corresponding plain text space P, cipher-text space C and key space Zn can be defined as P = C = Zn respectively to satisfy the Equ (3.4) and (3.5).

$$K := (N, e, d)|ed \equiv 1mod\phi(n) \qquad \text{(Equ:3.4)}$$

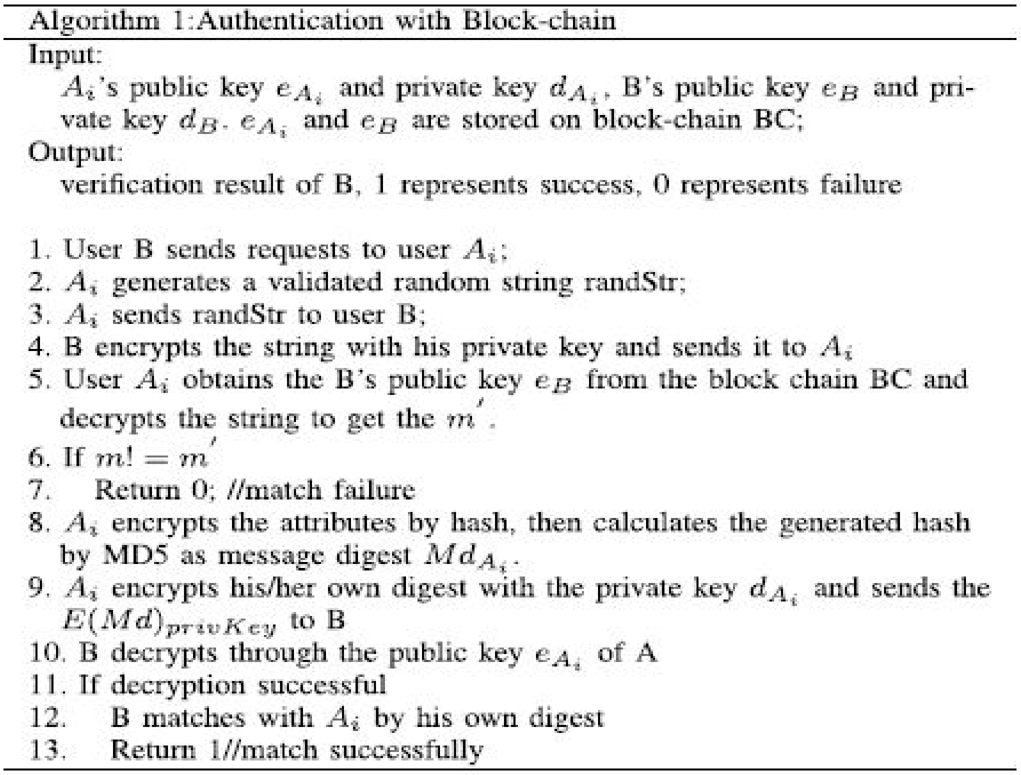
$$\phi(n) = p^r(p-1)(q-1) \qquad \text{(Equ:3.5)}$$

we get private Key e and public Key d respectively. The encryption e_k(m) of plain text m $\in$ P Zn $\rightarrow$ Zn as shown in Equ (3.6).

$$e_k(m) := m^e modn \qquad \text{(Equ:3.6)}$$

3.2.2 Block-chain algorithm

We implement authentication between social circle and user with the Authentication with Block-chain algorithm. Assume that A_i belongs to the user set U(A1, A2, ... , An) and B is the candidate user node to be added into that user set. We calculate the intimacy between user B and each user in social circle.



```
Algorithm 1:Authentication with Block-chain
Input:
   $A_i$'s public key $e_{A_i}$ and private key $d_{A_i}$, B's public key $e_B$ and pri-
   vate key $d_B$. $e_{A_i}$ and $e_B$ are stored on block-chain BC;
Output:
   verification result of B, 1 represents success, 0 represents failure

1. User B sends requests to user $A_i$;
2. $A_i$ generates a validated random string randStr;
3. $A_i$ sends randStr to user B;
4. B encrypts the string with his private key and sends it to $A_i$
5. User $A_i$ obtains the B's public key $e_B$ from the block chain BC and
   decrypts the string to get the $m'$.
6. If $m! = m'$
7.     Return 0; //match failure
8. $A_i$ encrypts the attributes by hash, then calculates the generated hash
   by MD5 as message digest $Md_{A_i}$.
9. $A_i$ encrypts his/her own digest with the private key $d_{A_i}$ and sends the
   $E(Md)_{privKey}$ to B
10. B decrypts through the public key $e_{A_i}$ of A
11. If decryption successful
12.     B matches with $A_i$ by his own digest
13.     Return 1//match successfully
```

Fig. 3.3: Block chain algorithm

First, users create their own public and private key pairs, and use their own block chain system to store the public key on the block chain nodes. A block chain consists of two nodes forming a private block chain. A certain user keeps his private key, and his public key is stored in the established block chain binding with address. System gives users permission of the smart contract in block chain to ensure that users can access the public key of the other users through address.

Second, user A_i needs to encrypt his relationship attributes before sending his information to B. He needs to prepare a relational attribute set, his own private key, his block chain address, and the public key. Suppose that the attribute set is $S^{Ai}D$ (s_1,s_2, s_3,....,s_n), we first compute the hash value H(S^{Ai})=(hs_1, hs_2,hs_3,...., hs_n) of the attributes S^{Ai} . Because the generated hash string is 64 bit, and the length is slight longer, the time overhead required for matching is relatively large, and the security of one time hash encryption is slightly worse at the same time. So we encrypt it twice on basis of H(S^{Ai}) and get MD5 digest: Md^{Ai}=(md(hs_1), md(hs_2), md(hs_3)....,md(hs_n)). Then, A_i encrypts its own MD5 digest with its own private key and sends the result E(Md)privKey to B.

Finally, B hashes his relational attributes before it is matched, and produces its own MD5 digest. After receiving the information sent by A_i, B decrypts the signature of A_i by decrypting it with A_i's public key. When the signature is true, B matches the A_i relationship digest with his own relationship digest. After the success of the match, user B and user A_i have higher intimacy. If the match fails, because A_i does not send specific information about relationship, B does not get the relationship of A_i and the privacy of user A_i can be protected. We calculate similarity Sim(A_i, B) between A_i and B according to digests. They can be regarded as two vectors and the similarity between A_i and B can be calculated as following Equ (3.7).

$$Sim(A_i, B) = \frac{Md^{Ai} * Md^{B}}{|Md^{Ai}| * |Md^{B}|} \qquad \text{(Equ:3.7)}$$

If the similarity is high enough, B and A_i would have high intimacy. Otherwise, the closeness of B and A_i does not meet the requirements. Because B receives only a hash value of the properties of A_i, the digest is encrypted twice, and the hash cipher is not invertible, thus user B doesn't know the text of the properties of user A_i.

3.3 Text encryption based on content recommendation

To facilitate content recommendation, the server generates a pair of public key e_s and private key d_s, which contributes to users to send their own message to server protectively. The same as the user's public key, the public key of the server and other relevant information of the public key, are stored in the block chain. Thus, when the user encrypts the information, the correct public key can be obtained directly from the block chain, thereby avoiding the threat that a malicious user sends the false public key and decrypts the information.

This process requires two steps: calculation and encryption of recommend content. Encryption algorithm is used twice. The first time is that before calculation of a certain user interest when the server needs to acquire some information about his own profile and posted message of other user. This kind of information needs to be encrypted. The second time is the encryption of the recommended messages.

In the first encryption process, when message M is sent to the system, the user obtains the corresponding system's public key e_s from the block chain. Then, user needs to encrypt M to obtain cipher text and send the cipher text to the system. After the system receives the cipher text, the system uses its own private key d_s to decrypt it.

3.3.1 Message Content Recommendation Algorithm

As for the problem about fulfilling the willing that users like to connect with similar users, we propose a personalized recommendation method named Message Content Recommendation Algorithm (MCRA) applied inside each of the communities, which considers the semantic analysis and statistical analysis together to recommend messages to centric user. In the original MCRA algorithm, semantic analysis and statistical analysis were taken into consideration to recommend messages to the central user. This process requires to calculate P(M|u) which is the semantic interest of user u for message M, and K(M|u) which is the information interest statistic of user u for message M.

Message Semantic Interest P(M|u): The Message Semantic Interest of user is defined as the possibility that User u uses Message M to express his opinion in terms of semantics.

Message Statistical Interest K(M|u). The Message Statistical Interest of user is defined as the possibility that User u uses Message M to express his opinion in terms of statistic.

To compute P(M|u), we use probabilistic method to describe the relationships between users and messages. The message is represented by the bag of words model. The idea is that one document exhibits multiple topics, and a topic is made of several words in the form of possibility. P(M|u) should be the possibility of each word. We set the maximum value of P(w|u) in all the words in the message to represent P(M|u) of the whole message in MCRA to avoid the problem of decreasing product value as the message length increases. The possibility of the word "w_i" issued by the user P(w_i |u) is defined as Equ (3.8), where T represents the topic, and T is the collection of topics U.

$$P(w_i|u) = \sum_{tT} P(t|u)P(w_i|t) \qquad \text{(Equ:3.8)}$$

In order to compute P(w_i |u), we use Latent Dirichlet Allocation (LDA) to train the model to achieve the user topic possibility distribution P(T|u) and the subject word possibility distribution P(V|T). The user topic possibility distribution P(T|u) is a vector whose elements are possibilities that a target user u is interested in each topic in topic collection T.

$$P(T|u) = P(t_0|u), P(t_1|u), ..., P(t_n|u) \qquad \text{(Equ:3.9)}$$

In Equ. 3.9, t_i is the i^{th} topic in T, and P(t_i | u) denotes the possibility that in what degree the messages posted by target user u match the topic t_i. The possibility between Topic t_i and term v_j is p(v_j | t_i), and the possibility of word distribution of Topic t_i is shown as Equ. 3.10.

$$P(V|t_i) = P(v_0|t_i), P(v_1|t_i), ..., P(v_{m0}|t_i) \qquad \text{(Equ:3.10)}$$

The Topic-Word possibility distribution P(V | T) is defined as follows(Equ. 3.11).Each topic contains numbers of words. And there are numbers of topics in topic collection.

$$P(V|T) = \begin{vmatrix} P(v_0|t_0) & ... & P(v_m|t_0) \\ ... & ... & ... \\ P(v_0|tm) & ... & P(v_m|t_m) \end{vmatrix} \qquad \text{(Equ:3.11)}$$

$P(M|u)$ could be calculated according to $P(w|u)$ of each word in a message. Generally thinking $P(M|u)$ should be the product of the possibility of each word. However, the length of messages are not quite same, thus leads to an unfair situation because the longer a message is, the smaller the product value tends to be. To avoid this problem, the highest value of $P(w|u)$ among all the values of words in the message will be set as $P(M|u)$ for the whole message in MCRA, as shown as Equ.3.12.

$$P(M|u) = maxP(w_0|u), P(w_1|u), ..., P(w_n|u) \qquad \text{(Equ:3.12)}$$

$K(M|u)$ denotes the importance of Message M to User u. To compute $K(M|u)$, the importance degree of each word to the user, i.e. $K(w|u)$, needs to be achieved.$K(M|u)$ would refer to user-word weight vector formed as shown as Eq. 3.13, where w(v_n, u) is the weight of v_n in the message collection of target user u.

$$K_u = w(v_0, u), w(v_1, u), ..., w(v_n, u); v_0, v_1, ... \in V_c \qquad \text{(Equ:3.13)}$$

In calculating K(M|u), we first need to compute the user word weight vector K(u), whose constituent elements are the weight w(v_i, u) of the message v_nx in the message set of the target user u. The policy for calculating K(M|u) is similar to that of P(M|u), such as Formula (3.14). If the candidate message contains no text in the content file of the target user, the similarity will be set to the minimum weight in K(u).

$$K(M|u) = \begin{cases} min(w(w_i), v_i \in V_c, \exists w_i \notin V_{targetusr} \\ max\{K(w_0|u), ..., K(w_n|u)\}, \exists w_i \in V_{targetusr} \end{cases} \quad \text{(Equ:3.14)}$$

In Equ.3.14, M = { w_1, w_2, . . . , w_m }, $K_u(w_n)$ is the interest degree that User u likes Word w_n. w(v_i, u) is the interest degree that User u likes to use Word v_i in his own file. Ku(W) is the interest degree that User u tends to post candidate message K.

After calculating the semantic interest P(M|u) and the information interest statistics K(M|u), the two parameters need to be considered comprehensively. We use the idea of weighted mean to introduce the concept of information score Score(M|u), and the information score is shown in Equ (3.15). After scoring, select the highest score messages M_{max} and send it to the user.

When Messages M_{max} are sent to the user, the system obtains the corresponding user's public key from the block chain and encrypts it to obtain c_m. Then, the system encrypts the user attributes using the system private key d_s and obtains the encrypted file c_{att} , which is sent to the user along with the c_m. The user can obtain the system public key e_s from the block chain, and then decrypt the c_{att} information sent by the system. If the decryption results correspond to their own attributes, then the information is sent by the system. The user then decrypts the cm using his private key, thereby preventing attackers from recommending junk files to the user and preventing attackers from stealing information.

$$Score(M|u) = \frac{\alpha P(M|u) + \beta K(M|u)}{2} \quad \text{(Equ:3.15)}$$

3.4 Performance evaluation

Social circle extension

While extending the community, there are two parameters need to be estimated, which are parameter k in KSM algorithm and threshold δ in CSE algorithm. First, k is assigned as a certain value and δ varies. Then, δ is set constant, and k changes.

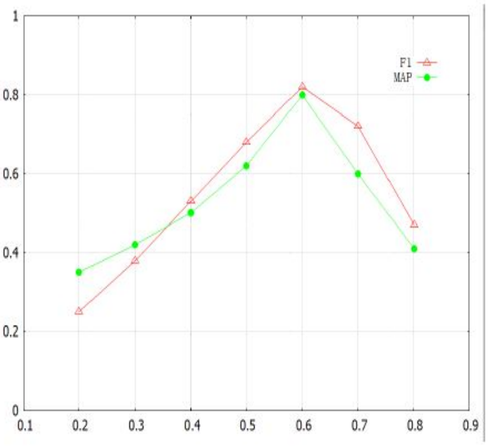


Fig. 3.4: Result of changing δ, when k=4

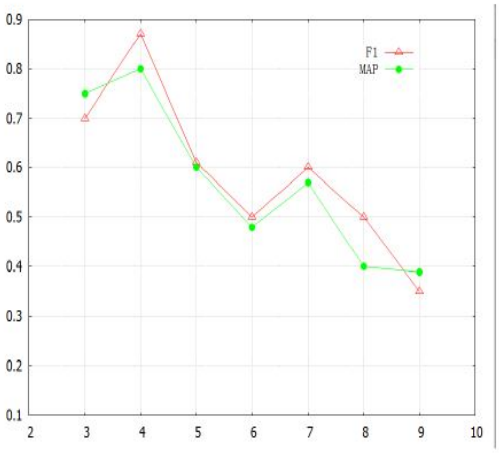


Fig. 3.5: Result of changing k when δ =0.6

From what can be seen in the result, when δ = 0.6 and k = 4, the best performance of the algorithm can be achieved.

- The authentication with block-chain is introduced to verify the validity of the user identity. At the same time, acknowledged and non-modifiable nature of the block chain is used to store the generated public key on the block chain to bind the block chain address to the public key in order to avoid other people's public key forgery and tampering. Identity can be confirmed only by authenticated users in order to avoid malicious users attacking legitimate users. Also, a user has an address corresponding to the public key, which avoids Sybil attacks to a certain extent.
- Meanwhile, to prevent honest but curious users take advantage of their interests to get interest from other users, relation information is not directly transferred in plain text when matching.A vector with relations is constructed, and make twice hash encryption for relation attributes. Because of the irreversibility of hash encryption, the user only gets the MD5 value of the property so that mismatched users cannot obtain the user relationship attributes. Only same attributes can be obtained for same MD5 to match.

- Using MD5, the dimension of the string after the first hash encryption can be reduced, then not only the twice hash make encryption more secure to guarantee the uniqueness of attribute encryption, but the number of bits is small and the matching process can increase matching efficiency.
- The technology of digital signature: in order to guarantee the identity of the user who sends interest, the user should encrypt the data packets with his private key.

The comparison of social circle

To verify the efficiency of the expansion, CMCR is compared with RSCM algorithm and K-means algorithm(Fig 3.6).

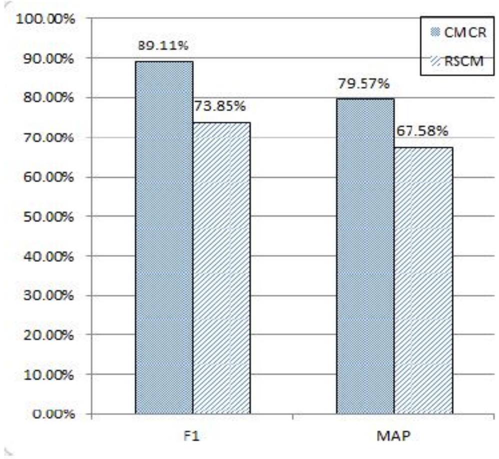


Fig. 3.6: Comparison of CMCR and RSCM.

When the social circle is expanded, K-means algorithm directly uses the number of previously detected communities as the initial number of centers. Moreover, K-means algorithm puts each node into a community. This algorithm can produce satisfactory results in detecting communities,compared with the RSCM algorithm and K-means algorithm.

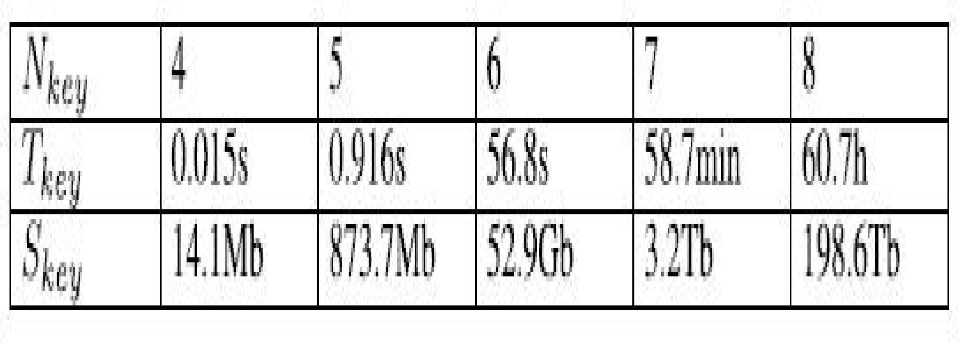
Security testing

Hash encryption algorithm is very effective. There are mainly two common method attacks on the hash encryption: find collision method and exhaustive method. The first method is finding collision. Different strings would lead to same hash values when a collision occurs. Therefore, the attacker could crack the MD5 by obtaining the same hash value with the one using in the encryption process.But there is no need to worry about this situation, since there is no effective way to find collision method for MD5 and SHA1.So far, the order of magnitude of the most effective crack algorithm to solve MD5 is 269. But this situation is still limited to theoretical analysis. In fact, 269 is still an impossible number for practical application.

Another method attacking the hash encryption method is exhaustive. For some simple password, this method is very efficient and easily implemented, for example, “123456” and “000000”. Because the scanning scope of the exhaustive method is often a single character set, interval with the law, or the combination of the words in the dictionary, so exhaustive method is difficult to work in the cases of complex passwords.The number of characters that may appear at each bit in the hash encryption method is Num_{possi}, the number of bits N_{key} and the time required to crack a password is T_{bit} . The cracked time T_{key} is defined as:

$$T_{key} = T_{bit} * Num_{possi}^{Nkey} \qquad \text{(Equ:3.16)}$$

The required storage space S_{key} can be calculated in a similar way. Assuming that Num_{possi} is 62 and T_{bit} is 1ns. It is almost impossible to break hash encryption with the ordinary exhaustive method.As the number of bits N_{key} increases, the cracked time T_{key} and the required storage space S_{key} are shown in Fig 3.6 .



N_{key}	4	5	6	7	8
T_{key}	0.015s	0.916s	56.8s	58.7min	60.7h
S_{key}	14.1Mb	873.7Mb	52.9Gb	3.2Tb	198.6Tb

Fig. 3.7: The time and storage of exhaustion.

The violent crack method can be simplified using a method called a rainbow table. Rainbow table is a precomputed table for breaking the hash value of the password, used in the inverse operation of the encryption hashing function. The rainbow tables are often used to recover the fixed length plain text passwords consisting of the characters in finite set.

In brief, this is an effective method cracking of some particular algorithms, especially the asymmetric algorithm, such as MD5 algorithm. Ignoring the time required for the query, the larger the table is, the lower the cost of cracking is. However, for other crack methods, such as collisions, the effect of cracking would be poor. Especially for variable length keys and other modern advanced algorithms, the effect will be greatly reduced. The use of secondary encryption method, guarantees that the possibility of being cracked is very small using the rainbow table

Based on the analysis, the brute force is still the main crack hash encryption method. So in order to reinforce the security of encrypted information, we use multiple hash methods. The method is called multi-hash that uses multiple encryptions on the information with the hash method through user-defined Key. If the Key is complicated enough, it's very difficult to crack in exhaustion method.

$$R = MD5(SHA1(S)) \tag{Equ:3.17}$$

Even if S is simple, it is still difficult to computing all possibilities in a reasonable time cracking the hash encryption and MD5. This method further ensures the security of the data.

Conclusion

Online social networks have gained huge popularity in today's world. People tend to use multiple networking sites such as Facebook, Twitter, and LinkedIn. Social networks have become an inevitable part in our daily life. Social network data contains sensitive and confidential information about users. Thus, in its original form, the exchange of these data may affect the privacy of individuals. Based on CMCR, a protocol and authentication with Block-chain algorithm to protect the user privacy information in the social community detection is proposed here. Two problems related to large scale mobile social network analysis are discussed. The first one is overlapping community detection and the other one is personalized content recommendation in communities. Considering user with higher closeness, we use authentication mechanism based on the block-chain, and encrypt the relationship with Hash function for better security. Then, we use the text encryption protocol in the text recommendation process to ensure the security of information. The algorithm is verified by contrast experiments, which show that the proposed algorithm can outperform the baseline algorithms in some common criteria.

REFERENCES

[1] R. Yu et al., “Communities mining and recommendation for large-scale mobile social networks,” in Proc. Int. Conf. Wireless Algorithms, Syst., Appl., Cham, Switzerland, 2017, pp. 200-205.

[2] J. Yang and J. Leskovec, “Overlapping community detection at scale: A non negative matrix factorization approach,” in Proc. ACM 6th Int. Conf. Web Search Data Mining, Rome, Italy, 2013, pp. 587-596.

[3] W. Dong, V. Dave, L. Qiu, and Y. Zhang, ”secure friend discovery in mobile social networks,” in proc. IEEE INFOCOM, Apr. 2011.

[4] S. Fortunato and C. Castellano, ”Community structure in graphs,” in Computational Complexity. New York, NY, USA: Springer-Verlag,2012, pp. 490-512.

[5] J. Lei, L. En-tao, and W. Guo-jun,”Privacy preserving friend matching mechanism in mobile social networks,” J Chin. Comput. Syst., vol. 37, no. 9, pp. 1980-1985, 2016.

[6] M. Nofer et al., “Blockchain,” Bus. Inf. Syst. Eng., vol. 59, no. 3, pp. 1–5, 2017.

[7] S. Nakamoto, “Bitcoin: A peer-to-peer electronic cash system,” [Online]. Available: https://bitcoin.org/bitcoin.pdf

[8] D. Fu and L. Fang, “ Block chain-based trusted computing in social network,” in Proc. IEEE Int. Conf. Comput. Commun., Oct. 2017, pp. 19-22.

[9] R. Ali and A.K. Pal, “ A secure and robust three factor authentication scheme using RSA cryptosystem,” Int. J. Bus. Data Commun. Netw., vol. 13, no. 1, pp. 74-84, 2017.