

TODO

- Undergraduate Research Opportunities -

Dragoş Alin Rotaru

Mentor: Ruxandra F. Olimid

Universitatea din Bucureşti

ROMANIA

`r.dragos0@gmail.com`

Abstract. TODO - abstract

Cuprins

TODO	1
<i>Dragoș Alin Rotaru Mentor: Ruxandra F. Olimid</i>	
1 Introducere	3
1.1 Istoric	3
1.2 Motivatie	3
1.3 Structura	3
1.4 Securitatea Teoretica a Informatiei	3
2 Scheme de partajare	4
2.1 Istoric	4
2.2 Schema unanimă	5
2.3 Schema Shamir	5
2.4 Schema Ito, Saito și Nishizeki	6
3 Sisteme de stocare de lunga durata	6
3.1 Criptare VS scheme de partajare	8
4 Sisteme de stocare de lungă durată bazate pe scheme de partajare ...	9
4.1 PASIS	9
4.2 GridSharing	10
4.3 POTSHARDS	12
5 Sistemul de stocare Alouneh et al.	13
6 Rezultate obținute	16
6.1 Vulnerabilități evidențiate	16
6.2 Implementare și rezultate practice	18
6.3 Publicarea rezultatelor	25

1 Introducere

1.1 Istoric

Termenul de criptografie este definit in dictionarul Oxford ca fiind "arta de a scrie si a rezolva coduri". Criptografia moderna s-a desprins de cea clasica in jurul anilor '80, motivand implementarea rigurozitatii matematice pentru definirea constructiilor criptografice. Asta pentru ca in anii anteriori, experienta a dovedit nesiguranta metodelor de criptare, criptanaliza lor fiind uneori triviala (cifrul lui Cezar, Vigenere [1], [2]) sau uneori atinsa cu ceva mai mult efort precum Enigma si alte metode din cel de-al doilea razboi mondial. ??

Criptografia moderna se gaseste pretutindeni in viata de zi cu zi de la ATM-uri, cartele telefonice la semnaturi digitale, protocoale de autentificare, licitatii electronice sau bani digitali, luand amploare o data cu aparitia sistemelor cu cheie publica. O definitie potrivita ar fi "studiul stiintific al tehnicilor pentru a securiza informatia digitala, tranzactiile si calculul distribuit.". [3]

1.2 Motivatie

TODO nu are legătură: faptul ca sunt scheme cuantice nu are impact asupra structurii de acces; e ok sa menționezi, dar atunci faci asta în secțiunea anterioară, când poți să adaugi și alte modalități de definire: scheme bazate pe latici, scheme bazate pe perechi biliniare, scheme cuantice, etc. Dezavantajul schemelor generale de partajare este dimensiunea componentelor, exponentiala in functie de numarul de participanti. [4] De asemenea, s-au dezvoltat scheme pentru modele de calcul unconventional, cum ar fi cel cuantic. [5]

1.3 Structura

TODO

1.4 Securitatea Teoretica a Informatiei

In cazul unor criptosisteme acestea nu pot fi compromise chiar daca adversarul dispune de o putere computationala nelimitata. Cateva exemple de criptosisteme care garanteaza securitatea teoretica-informationala sunt: schemele de partajare, unele protocoale multi-party computation, preluarea intr-un mod sigur(securizat?) informatii de la baze de date. Securitatea teoretica vine insa cu un cost: efortul computational depus este mult mai mare decat in cazul schemelor care nu garanteaza securitatea teoretica (se bazeaza pe dificultatea computationala unor probleme cunoscute). [6]

2 Scheme de partajare

O schemă de partajare constă în distribuirea unei informații secrete \mathcal{S} la mai mulți participanți $\mathcal{P} = \{P_1, \dots, P_n\}$ astfel încât oricare mulțime de participanți predefinită ca făcând parte dintr-o structură de acces pe care o vom denumi \mathcal{A} să poată reconstitui secretul \mathcal{S} . Formal, o schemă de partajare este reprezentată de o pereche de algoritmi (Gen, Rec):

- $Gen(\mathcal{S}, m)$ este un algoritm care primește la intrare un secret \mathcal{S} și un număr întreg m și întoarce un set de componente s_1, s_2, \dots, s_m .
- $Rec(s_{i1}, s_{i2}, \dots, s_{iq})$ este un algoritm care primește ca parametri de intrare o mulțime de componente și întoarce \mathcal{S} dacă mulțimea $\{P_{i1}, P_{i2}, \dots, P_{iq}\} \in \mathcal{A}$.

Majoritatea schemelor constau în mai multe etape precum:

- *Inițializare.* Presupune inițializarea variabilelor de mediu necesare.
- *Generare.* O entitate autorizată (numită dealer) \mathcal{D} folosește algoritmul Gen pentru a genera componentele.
- *Distribuire.* Componentele sunt trimise participanților cu ajutorul unui mijloc de comunicare sigur, fără ca acestea să fie vizibile unui atacator.
- *Reconstrucție.* Dându-se o mulțime de componente, se folosește algoritmul Rec pentru a recupera secretul \mathcal{S} .

Schemele de partajare se clasifică în funcție de cantitatea de informație secretă pe care o pot obține persoanele care nu fac parte din \mathcal{A} [7]:

- *Sisteme perfecte de partajare:* componentele nu oferă nici o informație teoretică despre \mathcal{S} indiferent de resursele computaționale.
- *Sisteme statistic sigure:* o fracțiune de informație este dezvaluită despre \mathcal{S} independent de puterea computațională a adversarului.
- *Sisteme computațional-sigure de partajare:* se bazează pe faptul că reconstituirea lui \mathcal{S} se reduce la o problemă *dificilă* (spre exemplu problema Diffie-Hellman [8]) în lipsa unor informații oferite doar grupului de acces \mathcal{A} .

În continuare vom prezenta câteva sisteme perfecte de partajare utilizate în cadrul unor arhitecturi pentru stocarea fișierelor pe o durată îndelungată.

2.1 Istoric

Primele scheme de partajare au fost dezvoltate independent de Shamir și Blakley în 1979 [9, 10].

Denumite și scheme majoritare (k, n) , acestea rezolvau cazul în care oricare grup de participanți cu un număr mai mare sau egal decât k (mărimea pragului) poate reconstitui secretul \mathcal{S} din componentele primite de la dealer. Dacă schema este

perfect *sigură* atunci oricare grup cu un număr de participanți mai mic decât k nu obține vreo informație despre \mathcal{S} .

Schemele majoritare (spre exemplu schema Shamir) sunt insuficiente pentru a permite partajarea pentru anumite structuri de acces. Considerăm cazul în care vrem să partajăm un secret între 4 participanți: P_1, P_2, P_3, P_4 astfel încât $\{P_1, P_2\}$ și $\{P_3, P_4\}$ să fie singurele mulțimi autorizate pentru reconstrucția secretului \mathcal{S} (i.e. $\mathcal{A} = \{\{P_1, P_2\}, \{P_3, P_4\}\}$). În mod evident, problema nu poate fi rezolvată cu o structură de acces de tip prag: anumite mulțimi de 2 participanți trebuie să poată reconstrui secretul ($\{P_1, P_2\}, \{P_3, P_4\}$), în timp ce altele nu ($\{P_1, P_3\}, \{P_1, P_4\}, \{P_2, P_3\}, \{P_2, P_4\}$)

Astfel de scheme de partajare pentru structuri de acces generale au fost dezvoltate de Ito, Saito și Nishizeki, realizând o generalizare a schemei Shamir [11]. Benaloh și Leichter au demonstrat că schemele de partajare de tip prag nu pot fi folosite pe structuri general monotone (familie de submulțimi ale lui \mathcal{P} cu proprietatea că dacă $A \in \mathcal{A}$ și $A \subset A'$, atunci $A' \in \mathcal{A}$) și obțin o construcție mai eficientă ca Ito et. al din punct de vedere al numărului de componente distribuite participanților [12].

2.2 Schema unanimă

Presupunând că vrem să împărțim un secret \mathcal{S} la n participanți astfel încât \mathcal{S} să poată fi recuperat doar dacă toți cei n participanți își combină componentele pe care le dețin. Metoda este echivalentă cu o schemă (n, n) majoritară. Un exemplu este schema introdusă de Karin, Greene și Hellman (Fig.1) [13].

2.3 Schema Shamir

Schema Shamir oferă mai multă flexibilitate decât schema unanimă prin faptul că oricare k (sau mai mulți) participanți din cei n pot recupera \mathcal{S} , însă mai puțin de k participanți nu obțin nicio informație despre \mathcal{S} . Schema Shamir este deci o schemă (k, n) majoritară.

Intuitiv, având k puncte în plan (x_i, y_i) , $x_i \neq x_j$ $i, j \in \{1, 2, \dots, k\} \forall i \neq j$, există o curbă polinomială unică care trece prin ele. În schimb, pentru a defini o curbă polinomială de grad k care trece prin $k - 1$ puncte date, există o infinitate de soluții. Evident, orice submulțime de valori s_i de mărime egală cu k este suficientă și necesară pentru a reconstrui polinomul f . După interpolarea componentelor deținute de cel puțin k dintre participanți, secretul \mathcal{S} se determină ca fiind $f(0)$ (Fig. 2) [10].

Pentru un atacator care deține chiar și $k - 1$ valori s_i , acesta nu determină nimic despre \mathcal{S} , spațiul de soluții posibile fiind identic față de situația în care nu reușește să obțină vreo componentă.

Inițializare:

- Fie $S \in Z_q$ unde $q > 1$ și q prim;
- Fie n numărul de participanți;

Generare: Dealerul \mathcal{D} :

- Alege $n - 1$ valori aleatoare $s_i \leftarrow^R Z_p, i \in \{1, 2, \dots, n - 1\}$;
- $s_n = S + \sum_{i=1}^{n-1} s_i \pmod{q}$;

Distribuție: Dealerul \mathcal{D} :

- transmite în mod sigur participantului P_i componenta $s_i, i \in \{1, 2, \dots, n\}$;

Reconstrucție: Cei n participanți:

- Calculează $S = \sum_{i=1}^n s_i \pmod{q}$.

Fig. 1: Schema unanimă [13]

2.4 Schema Ito, Saito și Nishizeki

În continuare vom descrie modalitatea de distribuire a componentelor de la care au pornit Ito, Saito și Nishizeki pentru ca schema să aibă o structură de acces $\mathcal{A} \subseteq 2^P$ (submulțime a setului de participanți) monotona (i.e. $\forall A \in \mathcal{A}, A \subseteq A' \Rightarrow A' \in \mathcal{A}$). [11] Folosind construcția unei scheme majoritare (k, n) autorii au reușit să descrie elementele din \mathcal{A} folosind rezultatul unei reuniuni de mulțimi de componente cu un număr de elemente mai mare sau egal decât k (Fig. 3) [11]. Notăția $x : Pr$, înseamnă că x are proprietatea Pr .

Dezavantajul acestei structuri este numărul de componente necesar pentru o structură de acces oarecare \mathcal{A} . Un mod simplu de construire al funcției *Assign* este următorul:

- Fie r numărul de elemente corespunzător mulțimii \mathcal{A} ;
- Se generează $\binom{n}{r}$ componente
-

3 Sisteme de stocare de lunga durata

În această secțiune vom arăta câteva întrebări ale schemelor de partajare. Considerăm cazul în care vrem să stocăm rapoarte medicale, imagini, docu-

Inițializare:

- Fie $S \in Z_q$ unde $q > 1$ și q prim;
- Fie n numărul de participanți a.i $q > n$;
- Fie k numărul minim de componente puse în comun pentru a determina pe S ;

Generare: Dealerul \mathcal{D} :

- Alege n valori distincte $x_i \leftarrow^R Z_q, i = 1, 2, \dots, n$;
- Alege $a_i \leftarrow^R Z_q, i \in \{1, 2, \dots, k-1\}, a_{k-1} \neq 0$;
- Construiește polinomul $f(x) = a_{k-1}x^{k-1} + a_{k-2}x^{k-2} + \dots + a_1x + S$;
- Calculează $s_i = f(x_i), i \in \{1, 2, \dots, n\}$;

Distribuție: Dealerul \mathcal{D} :

- Transmite participantului P_i componenta $s_i, i \in \{1, \dots, n-1\}$;

Reconstrucție: Orice mulțime cu dimensiunea k (sau mai mare) de participanți distincți P_1, P_2, \dots, P_k :

- Interpolează punctele s_i pentru a obține polinomul f :

$$f(x) = \sum_{i=1}^k s_i \prod_{1 \leq j \leq k, j \neq i} \frac{x - x_j}{x_i - x_j} \quad (1)$$

- Află secretul reconstruit $S = f(0)$.

Fig. 2: Schema Shamir [10]

mente clasificate pe un timp îndelungat într-un mediu electronic. Pe parcursul timpului, pot apărea în schimb, diverse probleme precum dezastre naturale, defectiunea unor componente hardware, eroare umană, etc [14]. Un sistem de stocare necesar nevoilor noastre trebuie să satisfacă cel puțin următoarele 3 condiții:

- Disponibilitatea: Informația trebuie să rămână accesibilă tot timpul, în ciuda erorilor de tip hardware.
- Integritatea: Abilitatea sistemului de a răspunde cererilor într-un mod care garantează corectitudinea lor.
- Confidentialitatea: O persoană care nu face parte din grupul de acces să nu obțină permisiunea de a afla informații de orice fel despre datele existente în sistem.

Inițializare:

- Fie q un număr prim $q, q > 1, z \in \mathbb{N}$ nenul și $\mathcal{C} = GF(p^z)$;
- Fie $S \in \mathcal{C}$ secretul;
- Fie structura de acces \mathcal{A} ;
- Fie n numărul de participanți;

Generare: Dealerul \mathcal{D} :

- Alege n valori distincte $x_i \leftarrow^R Z_q, i = 1, 2, \dots, n$;
- Alege $a_i \leftarrow^R \mathcal{C} \setminus \{0\}, i \in \{1, 2, \dots, k-1\}, a_{k-1} \neq 0$;
- Construiește polinomul $f(x) = a_{k-1}x^{k-1} + a_{k-2}x^{k-2} + \dots + a_1x + S$;
- Atribue $s_i = f(x_i) \ i \in \{1, 2, \dots, n\}$; Fie $Shares = \{s_1, \dots, s_n\}$;
- Alege $D_i \subseteq Shares \ 1 \leq i \leq n$;
- Alege funcția $Assign : P \rightarrow 2^Q$:
 - $Assign(P_i) = D_i \ 1 \leq i \leq n$
 - $\mathcal{A} = \left\{ Q \subseteq Shares : \left| \bigcup_{P_i \in Q} Assign(P_i) \right| \geq k \right\}$;

Distribuție: Dealerul \mathcal{D} :

- Transmite participantului P_i componenta $Assign(P_i), i \in 1, 2, \dots, n$;

Reconstrucție: Participanții din structura de acces \mathcal{A} :

- Procedenza identic ca in schema Shamir.

Fig. 3: Schema Ito, Saito, si Nishizeki [11]

3.1 Criptare VS scheme de partajare

Una dintre solutiile existente pentru a construi acest sistem ar putea fi criptarea datelor folosind o cheie inainte de inserarea lor in spatiul de stocare. In momentul in care un user autorizat doreste sa efectueze o citire a unor date, intrebuinteaza cheia potrivita pentru a le decripta. In practica exista algoritmi de criptare eficienti precum AES insă aceștia nu garantează confidențialitatea datelor în cazul în care avem de a face cu un adversar fara o limita computationala. Un dezavantaj al criptării este administrarea cheilor, standardele de securitatea schimbându-se in fiecare an. De fiecare data cand cheile sunt inlocuite atunci este necesara recriptarea datelor de pe fiecare baza de date. Cu cat disponibilitatea este mai mare - numarul de noduri duplicate crește- recriptarea lor devine o operatie costisitoare.

Majoritatea tehnicilor de criptare se bazează pe dificultatea factorizării unui număr sau cea a calculării logaritmului discret însă o dată cu posibila dezvoltare a calculatoarelor cuantice aceste probleme nu vor mai fi atât de dificile [15].

4 Sisteme de stocare de lungă durată bazate pe scheme de partajare

O alternativă la soluția cu criptare care asigură atât confidențialitate cât și redundanță necesară este întrebuințarea sistemelor de stocare de lungă durată bazate pe scheme de partajare [14, 16, 17].

4.1 PASIS

PASIS este o soluție pentru un sistem descentralizat care oferă beneficii precum securitate, redundanță a datelor și auto-întreținere. Structurile descentralizate împart informația la mai multe noduri folosind scheme de redundanță precum RAID **TODO (R... A... I... D...)** - mereu trebuie în paranteză denumirea completă, o singură dată, când se introduce o abreviere pentru a asigura performanța, scalabilitatea sistemului dar și integritatea datelor [18]. **TODO exact, câteva detalii despre RAID**

PASIS folosește schemele de partajare pentru a distribui informația nodurilor de stocare dintr-o rețea. Aceasta introduce agenți pe partea clientului pentru a scrie sau șterge date din noduri, dar și agenți pentru mentenanță. Componentele obținute **TODO în urma partajării unui** fișier sunt **TODO nu puse, ci stocate** în rețea cu ajutorul agenților (Fig. 4). Pe lângă conținutul brut al componentelor se adaugă metadate pentru a reține adresa nodului din rețea la care au fost **TODO stocate, nu trimise** dar și noua denumire cu care este salvată în rețea.

Considerând o schemă de partajare majoritară $p - m - n$ unde oricare din cei m participanți pot reconstitui fișierul, dar mai puțin de p nu obțin nicio informație dintr-un total de n componente. **TODO o schema de partajare majoritară are 2 parametrii, care sunt aici?! rescriere! atenție, să folosești notațiile de mai înainte: n = nr de participanți, etc.**

Atunci când un participant inițiază o cerere pentru a citi un fișier, agentul PASIS aflat local procedează după cum urmează:

- Caută numele celor n componente care alcătuiesc fișierul într-un serviciu care listează toate datele.
- Inițiază cereri de citire la cel puțin m din cele n noduri.
- În caz ca acesta nu primește cel puțin m răspunsuri se întoarce la pasul anterior încercând interogări la noduri diferite.

- Reconstituie fișierul obținut din cele m componente.

Operația de scriere este similară cu cea de citire, aceasta oprindu-se atunci când în cel puțin $n - m + 1$ noduri s-au stocat cu succes componente. În articol se menționează și compromisul de spațiu-timp **TODO denumirea consacrată în romana este compromis spațiu-timp, nu timp-spațiu** folosite de PASIS **TODO : extinde puțin ce înseamnă asta**. Autorii specifică soluții pentru auto mentenanța sistemului cu ajutorul resurselor umane prin monitorizarea periodică stării sistemului folosind log-uri sau ajustarea parametrilor din cadrul schemei de partajare.

TODO Imi place figura! Dar citeaza lucrarea - și în caption și în text - nu am vazut nicio citare la PASIS!

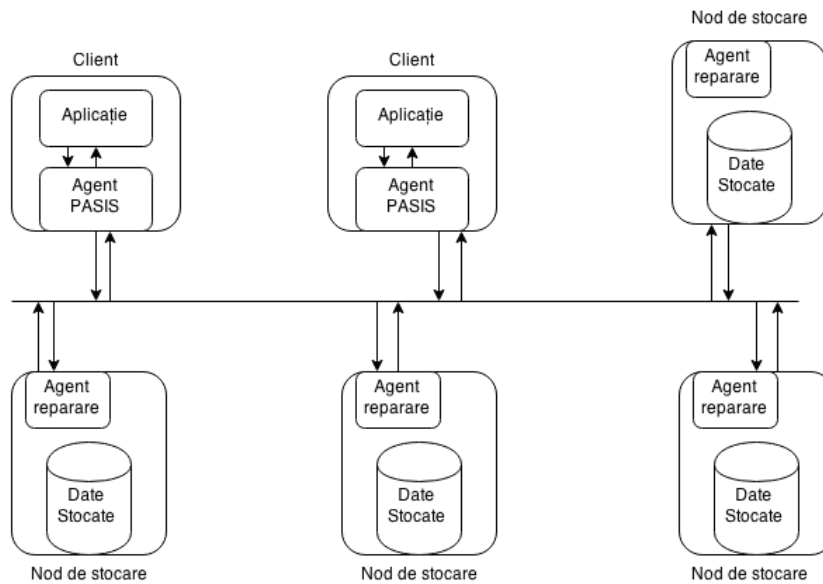


Fig. 4: Arhitectura PASIS cu 4 noduri și 2 clienți []

4.2 GridSharing

În 2005, Subbiah și Blough propun o nouă abordare pentru a construi un sistem de stocare securizat și tolerant la erori numit GridSharing [17].

Schema Shamir nu oferă siguranță în ceea ce privește detectarea sau actualizarea unor componente incorecte introduse de un atacator. Metoda cea mai des folosită

este determinarea validității componentelor prin utilizarea semnăturilor electronice.

Aceasta este realizată prin scheme de verificare non-interactive precum cea a lui Feldman împreună cu schema Shamir [19] **TODO trebuie reformulat: (1) leaga-l de paragraful anterior; (2) Feldman e o schema de sine statatoare, doar ca este construita pe baza schemei Shamir; cum se folosesc amandoua?**

Subbiah și Blough folosesc un sistem care înlocuiește schemele de verificare cu o schemă de partajare unanimă XOR (considerăm cazul $q = 2$ în Fig. 1) pentru a păstra securitatea construcției. În cazul detectării componentelor incorecte, este adoptată o strategie de tipul replicate-and-voting. Componentele sunt replicate pe un număr mare de servere astfel încât determinarea validității va fi stabilită în funcție de numărul de servere care le conțin.

Se identifică 3 tipuri de defecțiuni care pot apărea pe serverele unde sunt stocate datele:

- Abandonări: un server este *abandonat* dacă nu mai răspunde vreunui mesaj din rețea și s-a oprit din a mai efectua vreo operație.
- Bizantine: atunci când serverul nu respectă întotdeauna protocoalele inițiale iar componentele salvate local au fost compromise.
- Scurgeri de informații: serverul execută protocoalele corect dar e posibil ca un adversar să fi obținut componentele stocate.

Primele 2 modele definite mai sus sunt preluate din calculul cu sisteme distribuite. Cel de-al 3-lea model a fost introdus pentru a defini atacatorul care folosește vulnerabilitățile cu intenția de a *învața* din informații.

Arhitectura GridSharing constă în N servere unde cel mult c servere pot fi abandonate, b servere bizantine și l cu scurgeri de informații. Cele N pot fi aranjate într-un grid cu r linii și N/r coloane (considerăm pentru simplitate că $N \pmod{r} = 0$). Caracteristicile modelului bizantin și cel specific scurgerilor de informații permit dezvăluirea componentelor unui adversar de pe cel mult $l + b$ servere.

Exemplul 1. Notăm $\binom{x}{y}$ fiind combinații de x elemente grupate câte y . Considerăm ca împărțim un secret \mathcal{S} la 4 linii (participanți) astfel încât sistemul să permită 2 componente de tip b , 1 componentă de tip l și 20 servere. În cazul acesta vom folosi o schemă majoritară XOR $\left(\binom{4}{3}, \binom{4}{3}\right) = (4, 4)$.

Vom avea 4 componente, (s_1, s_2, s_3, s_4) a.î. $s_1 \oplus s_2 \oplus s_3 \oplus s_4 = \mathcal{S}$. Distribuirea se face în felul următor:

- Serverele situate pe prima linie primesc s_1
- Serverele situate pe a 2-a linie primesc s_2
- Serverele situate pe a 3-a linie primesc s_3

- Serverele situate pe a 4-a linie primesc s_4

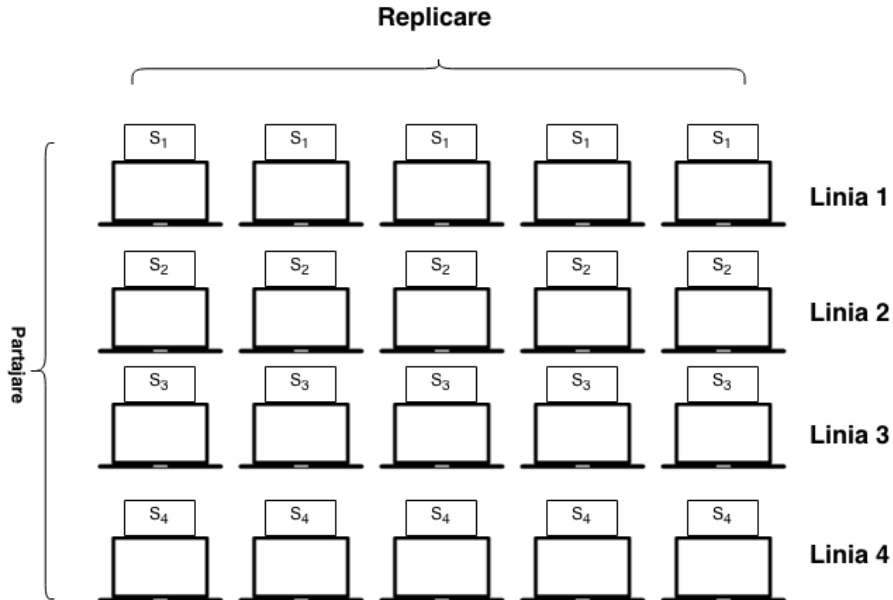


Fig. 5: GridSharing cu 4 linii, 20 servere dintre care 2 bizantine, 1 cu scurgeri de informații [17].

4.3 POTSHARDS

În 2007 este propus un nou sistem care combină caracteristicile PASIS și Grid-Sharing adăugând posibilitatea de migrarea a datelor la noduri noi: Protection Over Time, Securely Harboring And Reliably Distributing Stuff (POTSHARDS) [14]

De asemenea este introdusă o tehnică nouă de găsim a componentelor folosind pointeri aproximativi. Pentru a asigura confidențialitatea, autorii adoptă o schemă de partajare XOR unanimă, la fel ca în GridSharing. POTSHARDS consideră problema în care o persoană neautorizată încearcă să afle informații vulnerabile fără ca aceasta să fie nedetectată. Schemele existente precum PASIS și GridSharing nu îndeplineau această cerință dacă un atacator determină locația componentelor distribuite inițial. **TODO nu inteleg - reformulare, explica ce vrei sa zici!**

Soluția pe care o oferă această arhitectură este reconstruirea componentelor într-un mod securizat și folosirea semnăturilor algebrice pentru a asigura un grad ridicat de păstrare a integrității fișierelor [20]. POTSHARDS poate fi gândit ca o aplicație pe partea clientului care comunică cu o mulțime de noduri (arhive) independente.

Ca prim pas, POTSHARDS preprocesează fișierul într-un obiect, partajează obiectul în fragmente la care adaugă meta-date, numite *shards* (Fig. 6) [14]. Acestea sunt trimise apoi arhivelor independente, fiecare având propriul domeniu de securitate, localizate în *regiuni*. Pentru a reconstitui cu succes informația inițială, meta-datele shard-urilor conțin detalii despre structura pointerilor aproximativi, indicând regiunea în care se află următorul shard.

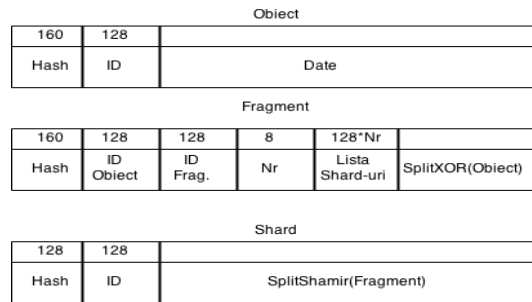


Fig. 6: Entități de date în POTSHARDS. *Nr* e numărul de shard-uri produse de un fragment. *SplitXOR* reprezintă o componentă rezultată în urma partajării unanime XOR. Analog *SplitShamir* reprezintă o componentă rezultată în urma partajării folosind schema Shamir. **TODO cite!**

Procesul de fragmentare a datelor este prezentat în Fig. 7.

Pentru ca reconstituirea unui fișier să fie fezabilă unui utilizator, acestuia îi este întoarsă o listă cu locațiile exacte shard-urilor corespunzătoare. Obținerea unui shard de către un atacator nu este folositoare, pentru a detecta următorul shard, un atac brut force constă în cereri multiple în zona indicată de pointerul aproximativ. Un astfel de atac nu va trece neobservat de POTSHARDS deoarece unul dintre scopurile sale este să stocheze datele într-un mod cât mai uniform distribuit **TODO spread**. [14]

5 Sistemul de stocare Alouneh et al.

Alouneh et al. propun un sistem pentru stocarea sigură a datelor un timp îndelungat folosind schema Shamir cu câteva modificări. Aceste schimbări se vor arăta esențiale mai târziu în menținerea securității [21].

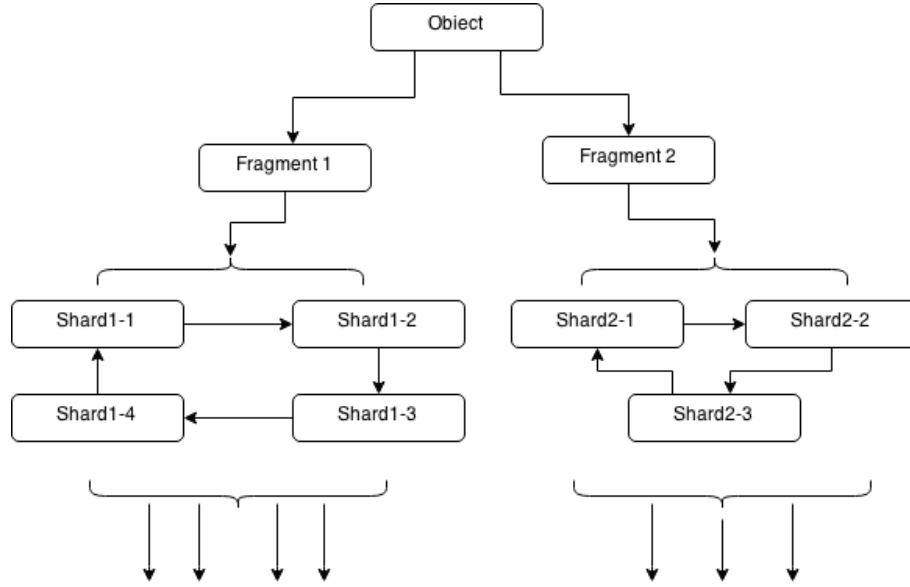


Fig. 7: Distribuirea unui obiect in POTSHARDS

Pentru stocarea unui fișier în sistem (abordând filozofia majorității sistemelor de operare - orice este un fișier), o aplicație client îl împarte în blocuri de octeți de lungime k . Pentru fiecare bloc, octeții devin coeficienții unui polinom f de grad k , componenta corespunzătoare participantului i va fi reprezentată de valoarea lui $f(i)$, $i = \{1, 2, \dots, n\}$. Menționăm că toate operațiile se vor efectua în $GF(256)$ modulo un polinom ireductibil (în implementarea sistemului, autorii folosesc $x^8 + x^5 + x^3 + x + 1$). Procedeu este descris în detaliu în Fig 8.

Exemplul 2. Vom exemplifica modul de calcul în $GF(256) \pmod{g(x)}$ unde $g(x) = x^8 + x^4 + x^3 + 1$. Luăm polinomul $f(x) = 10 + 15x$ corespunzător unui fișier format din octeții (în această ordine) 10 15.

$$\begin{aligned}
 f(01) \pmod{g(x)} &= 10 + 15 \pmod{g(x)} \\
 &= (x^4) + (x^4 + x^2 + 1) \pmod{g(x)} \\
 &= x^2 + 1 = 000000101_{(2)} \\
 &= 05_{(16)}
 \end{aligned} \tag{2}$$

$$\begin{aligned}
 f(02) \pmod{g(x)} &= 10 + 15 \cdot 02 \pmod{g(x)} \\
 &= (x^4) + (x^5 + x^3 + x) \pmod{g(x)} \\
 &= 00111010_{(2)} \\
 &= 3A_{(16)}
 \end{aligned} \tag{3}$$

Date de intrare: Un fișier binar \mathcal{S} ;

Date de ieșire: n fișiere binare distribuite la noduri din rețea;

Procesarea componentelor: Aplicația existentă pe partea clientului:

- Dacă \mathcal{S} nu are o lungime divizibilă cu k :
 - Concatenează la sfârșitul lui \mathcal{S} octeți până când $len(\mathcal{S}) \pmod k = 0$;
- Împarte \mathcal{S} în blocuri de lungime k ;
- Repetă pentru fiecare bloc B_t de lungime k :
 - Construiește polinomul $f(x) = B_{t_{k-1}}x^{k-1} + B_{t_{k-2}}x^{k-2} + \dots + B_{t_1}x + B_{t_0}$;
 - Calculează $f(i)$ pentru $1 \leq i \leq n$;

Distribuție: Aplicația la nivelul clientului:

- Distribuie componenta $f(i)$ nodului din rețea cu indicele i :

Fig. 8: Schema Alouneh et al. - Generare [21]

Reconstituirea unui fișier (Fig. 9) se realizează din orice mulțime de componente A de dimensiune minim k folosind interpolarea Lagrange (utilizată și în cadrul schemei Shamir):

$$f(x) = \sum_{i \in A} f(i) \prod_{j \in A, j \neq i} \frac{x - j}{i - j} \quad (4)$$

Exemplul 3. Vom exemplifica interpolarea conform ecuației 4 folosind componentele (2) și (3) calculate în exemplul anterior:

$$\begin{aligned} f(x) &= 05(x - 02)(01 - 02)^{-1} + 3A(x - 01)(02 - 01)^{-1} \\ &= 05(x - 02)03^{-1} + 3A(x - 01)03^{-1} \\ &= F6(05 + 3A)x + F6(05 \cdot 02 + 3A \cdot 01) \\ &= F6 \cdot 3F \cdot x + F6 \cdot 30 = 15x + 10 \end{aligned} \quad (5)$$

Noutatea arhitecturii constă în diminuarea redundanței componentelor la un factor de k , spre deosebire de sistemele descrise în Secțiunile 4.1, 4.2 și 4.3. Reducerea spațiului ocupat de componente este datorat înlocuirii coeficienților generați aleator din schema Shamir cu octeții din fișierul ce va fi partajat, fiecare componentă având nevoie de $1/k$ din dimensiunea fișierului \mathcal{S} .

Abordarea acestei metode deterministe conduce la insecuritatea sistemului, Alouneh et al. afirmând în mod eronat că securitatea este indusă în mod automat de schema Shamir.

TODO poti incerca sa adaugi si aici o imagine

- Date de intrare:** Cel puțin k componente provenite din noduri (distincte);
- Date de ieșire:** Fișierul binar original \mathcal{S} ;
- Reconstrucție:** Aplicația existentă pe partea clientului:
- Repetă pentru fiecare bloc al lui \mathcal{S} :
 - Calculează prin interpolare coeficienții lui $f(x) = B_{t_{k-1}}x^{k-1} + B_{t_{k-2}}x^{k-2} + \dots + B_{t_1} + B_{t_0}$
 - Reconstituie blocul B_t
 - Șterge octeții de la sfârșitul fișierului adăugați la generare.

Fig. 9: Schema Alouneh et al. - Reconstrucție [21]

6 Rezultate obținute

Împreună cu mentorul am analizat sistemul Alouneh et al. prezentat în Secțiunea 5 [21]. Am identificat erori majore ale acestuia și am implementat tot sistemul descris de autori pentru a demonstra practic, nu doar teoretic anumite greșeli pe care le vom evidenția în următoarele secțiuni.

6.1 Vulnerabilități evidențiate

Spre deosebire de schema Shamir, unde coeficienții (cu excepția termenului liber, care este egal cu secretul) sunt aleși într-un mod aleator uniform, sistemul propus de Alouneh et al. folosește o schemă modificată Shamir în care coeficienții sunt extrași din conținutul fișierelor originale. Alegerea este motivată de faptul că mulțimea componentelor și efortul computațional depus pentru generarea coeficienților se reduce la un factor de k ori, spre deosebire de schema Shamir.

Datorită determinismului, partajarea unui fișier de mai multe ori implică obținerea de componente identice. Determinismul duce la câteva atacuri simple în momentul în care un atacator obține informațiile stocate într-un nod, indiferent de mărimea pragului folosit în metoda de partajare. În acest sens, am evidențiat 2 atacuri simple în cazul în care componentele sunt calculate în ordine $(f(01), f(02), \dots)$:

- Detectarea tipului unui fișier
- Detectarea conținutului unui fișier

Deoarece Alouneh et al. nu menționează o metoda de padding, am indicat că un atac bazat pe felul în care se realizează completarea fișierului (padding) \mathcal{S} înainte de partajare sa poate fi fezabil, în condițiile în care s-a demonstrat că această alegere este esențială în păstrarea securității [22].

6.1.1 Detectarea tipului de fișier

În tehnologia informației, la începutul fiecărui fișier se află o secvență de octeți (denumită semnătură sau antet) cu rolul de a identifica tipul acestuia. Tabelul 1 indică 7 din cele mai uzuale antete.

Se consideră cazul în care partajarea unui fișier *pdf* se face cu ajutorul sistemului descris în Secțiunea 5, folosind $k \leq 4$. Polinomul corespunzător $f(x)$ va fi întotdeauna același. Presupunând ca numerotarea nodului i este aceeași, putem determina cu ușurință dacă este stocat un fișier *pdf* fără a lua în calcul conținutul fișierului. Acest atac este fezabil deoarece valoarea lui k este publică iar i poate să fie descoperit în momentul distribuirii.

Cu alte cuvinte, dacă un adversar obține controlul unui singur nod, bazându-se doar pe valoarea primei componente poate detecta tipul unui fișier.

Pentru a exemplifica, un adversar poate distinge cu probabilitate ridicată între fișierele *doc*, *gif*, *pdf*, *png*, *rar*, *wav* și *zip*. În Tabelul 2 avem generate componentele pentru $k = 2$ și $n = 5$. Dacă un adversar descoperă valoarea primului nod iar prima componentă este 14 atunci acesta știe cu certitudine că aceasta corespunde un fișier *gif*. Dacă obține accesul nodului 4 și citește valoarea 205 atunci știe că fișierul este de tipul *rar*. Dacă citește valoarea 27 de pe primul nod atunci știe că poate fi un *wav* sau *zip*; dar poate să distingă cele 2 fișiere cu probabilitate 1 dacă dezvăluie o singură valoare de pe celelalte noduri (2, 3, 4 sau 5) pentru că valorile sunt distincte.

Tabel 1: Semnături de fișiere

Tip de fișier	Primii 4 octeți
doc	D0 CF 11 E0
gif	47 49 46 38
pdf	25 50 44 46
png	89 50 4E 47
rar	52 61 72 21
wav	52 49 46 46
zip	50 4B 03 04

6.1.2 Detectarea de conținut

Multe documente urmează un anumit tipar precum contracte, chitanțe, bonuri fiscale sau curriculum vitae. Deoarece majoritatea conținutului rămâne neschimbat, există o probabilitate destul de mare ca multe componente să aibă aceeași

Tabel 2: Componentele primului bloc ($k = 2$)

Tip fisier	Nod 1 ($i = 1$)	Nod 2 ($i = 2$)	Nod 3 ($i = 3$)	Nod 4 ($i = 4$)	Nod 5 ($i = 5$)
doc	31	85	154	193	14
gif	14	213	156	120	49
pdf	117	133	213	126	46
png	217	41	121	210	130
rar	51	144	241	205	172
wav	27	192	137	109	36
zip	27	198	141	103	44

valoare. O dată ce un adversar reușește să determine componentele unui nod, poate determina prin analogie tipul de conținut al fișierului original.

Fișierele vulnerabile sunt cele care conțin o secvență de octeți periodică (imagini cu un pattern repetitiv) sau cele care au preponderent octeți nuli (valoarea componentelor asociată majorității blocurilor fiind 0). Spre deosebire de prima metodă, aceasta nu necesită compararea cu un al 2-lea fișier, tratând componentele de sine stătător. Multiple componente identice indică existența unor secvențe repetitive în conținutul fișierului.

6.2 Implementare și rezultate practice

Pentru a arăta aplicabilitatea rezultatelor în practică, am implementat propunerea descrisă în Secțiunea 5 și am testat pe câteva cazuri.

În cadrul implementării am folosit limbajul Python 3.0 sub sistemul de operare ArchLinux. Python este un limbaj high-level, permițând programatorilor să exprime concepte în mai puține linii de cod față de C++ sau Java și este disponibil sub licență open-source [23]. Pentru a realiza comunicarea între procese am folosit Cerealizer iar distribuția componentelor a fost generată grafic cu ajutorul pachetului Matplotlib [24,25]. De asemenea am folosit sistemul de versionare Git iar în prezent codul folosit este găzduit de GitHub [26,27].

Având în vedere că articolul original nu menționează o metodă de padding, am considerat o metodă standard pentru a completa octeții ultimului bloc: alipim la sfârșitul lui \mathcal{S} octeții 80 00 ... 00 00 până când lungimea ultimului bloc ajunge la k octeți.

Menționăm această metodă doar pentru completitudine, aceasta neafectând rezultatele, care consideră doar secvențele de octeți din antet sau de la începutul fișierului.

6.2.1 Detectarea tipului de fișier

Extindem analiza făcută în Secțiunea 6.1.1 asupra tipurilor de fișiere din Tabelul 1 pentru $k = 2$ și creștem valoarea indicelui i până când 2 componente devin egale. Fie f_l polinomul de gradul 1 asociat primului bloc al fișierului aflat pe linia l . Analog f_c polinomul de gradul 1 asociat primului bloc al fișierului aflat pe coloana c .

Tabelul 3 prezintă valoarea maximă a nodul i pentru care $f_l(i) \neq f_c(i)$. Valoarea -1 indică lipsă de coliziuni ale lui $f_c(x)$, $f_l(x)$ pentru i , $i = 1, 2, \dots, 255$ ($\nexists 1 \leq i \leq 255$ a.î. $f_l(i) = f_c(i)$)).

Deoarece pe diagonala principală toate componentele sunt identice pentru $k \leq 4$, poate fi ignorată. În Tabelul 3 observăm valoarea 0 pentru perechea (wav, zip) pentru că $f_{wav}(1) = f_{zip}(1) = 27$ (Tabelul 1).

Tabelele 4 și 5 prezintă rezultatele pentru $k = 3$ și $k = 4$. Pentru $k \geq 5$ este nevoie de un antet cu mai mult de 4 octeți.

Considerăm metoda de distribuire descrisă exact ca în Tabelul 8, și anume nodul de stocare i primește componenta $f(i)$. Dacă un atacator preia controlul nodului cu valoarea i acesta poate face distincția între tipul a 2 fișiere partajate cu probabilitate 1 dacă i e mai mic decât valoarea afișată în tabel.

Spre exemplu, un adversar care obține accesul unui singur nod de stocare, iar partajarea a fost făcută pentru $k = 2$ poate distinge cu probabilitate 1 între un fișier *doc* sau *pdf* dacă $i \leq 194$. Deoarece $n > 194$ nu se întâmplă în practică, acest atac funcționează. Faptul că multe valori sunt ridicate și majoritatea sunt -1 (Tabelul 4), indică un nivel scăzut de securitate al schemei. În cazul celor cu -1 un adversar va câștiga întotdeauna cu probabilitatea 1, indiferent de indicele nodului de stocare pe care îl obține.

Precizăm că acest atac funcționează doar dacă nodurile de stocare își păstrează indexul în cazul partajării repetate a.î. aplicația disponibilă pe partea de client calculează cele n valori $f(i_1), \dots, f(i_2)$ pentru i_1, \dots, i_n distincte și păstrează nodul j asociat lui i_j .

Tabel 3: Indicele maxim i a.î. componentele primului bloc sa fie distincte($k = 2$)

Tip Fișier	doc	gif	pdf	png	rar	wav	zip
doc	-	169	194	209	170	206	110
gif	169	-	133	137	75	-1	133
pdf	194	133	-	-1	115	151	133
png	209	137	-1	-	229	147	195
rar	170	75	115	229	-	-1	42
wav	206	-1	151	147	-1	-	0
zip	110	133	133	195	42	0	-

Tabel 4: Indicele maxim i a.î. componentele primului bloc sa fie distincte($k = 3$)

Tip Fişier	doc	gif	pdf	png	rar	wav	zip
doc	-	63	-1	-1	-1	-1	-1
gif	63	-	-1	-1	-1	-1	-1
pdf	-1	-1	-	164	-1	119	-1
png	-1	-1	164	-	143	122	129
rar	-1	-1	-1	143	-	143	-1
wav	-1	-1	119	122	143	-	172
zip	-1	-1	-1	129	-1	172	-

Tabel 5: Indicele maxim i a.î. componentele primului bloc sa fie distincte($k = 4$)

Tip Fişier	doc	gif	pdf	png	rar	wav	zip
doc	-	-1	38	95	1	95	98
gif	-1	-	-1	-1	167	-1	-1
pdf	38	-1	-	12	11	119	70
png	95	-1	12	-	243	95	148
rar	1	167	11	243	-	-1	94
wav	95	-1	119	95	-1	-	-1
zip	98	-1	70	148	94	-1	-

6.2.2 Detectarea de conţinut

Considerăm scenariul pentru detectarea conţinutului unui fişier, demonstrând practic cum un adversar poate să facă diferenţa între 2 componente stocate pe acelaşi nod aparţin unor documente similare. Atacul presupune accesul la un singur nod, indiferent de mărimea pragului k .

Pentru experiment am ales 3 fişiere *pdf* disponibile online la [?]:

- Europass Curriculum Vitae - BG, Bulgaria;
- Europass Curriculum Vitae - DK, Dannemark;
- Europass Mobility - RO, Romania.

Primele două fişiere reprezintă şablonul pentru un CV european în limba bulgară, respectiv daneză. Observăm că nu doar conţinutul şabloanelor diferă, dar şi limba în care au fost traduse. Cel de-al treilea fişier este complet diferit de primele două, fiind un document personal în limba română, folosit pentru înregistrarea cunoştinţelor dobândite într-o ţară europeană.

Pentru a arăta vulnerabilitatea sistemului Alouneh et al. [21], partajăm cele 3 fișiere folosind schema (2, 4). Experimentul reprezintă o implementare practică a metodei prezentate în Fig. 8.

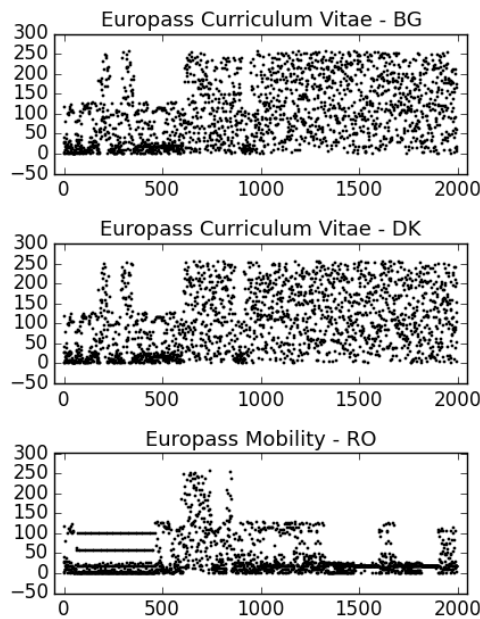


Fig. 10: Nod 1: Graficul componentelor pentru primele 2000 de blocuri

Fig. 10, 11, 12, 13 conțin 2000 de componente generate pentru cele 3 fișiere, stocate în baza de date corespunzătoare nodurilor 1, 2, 3 și 4. Presupunem că indicele i este fixat, polinoamele fiind evaluate în același nod i . Un adversar care obține componentele unui nod poate deduce cu ușurință graficul lor.

Considerăm că un adversar reușește să obțină componentele din Fig. 10. Poate deduce ușor că primele două fișiere au un număr mare de componente identice iar cel de-al 3-lea este diferit de primele.

Din cauza șablonului pe care îl respectă cele două CV-uri, un adversar are nevoie în cazul de față doar de primele 500 de componente pentru a constata similitudinea celor 2 cu o probabilitate ridicată. Diferența vizuală este evidentă între oricare din cele 4 grafice, deci un adversar poate reuși să afle informații despre fișierele partajate indiferent de nodul de stocare vulnerabil.

Considerăm în continuare un alt scenariu, în care partajăm o imagine pentru care informația se repetă. Fie imaginea din Fig. 14 partajată folosind schema

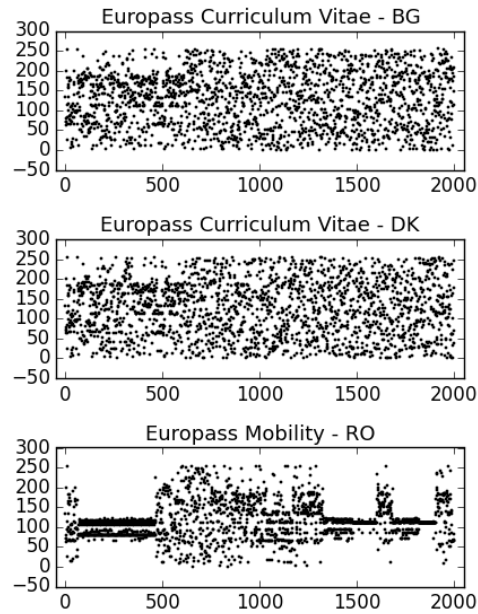


Fig. 11: Nod 2: Graficul componentelor pentru primele 2000 de blocuri

din Secțiunea 5 având 4 noduri iar informația stocată pe cel puțin 2 din ele pot reconstitui fișierul partajat inițial **TODO E mai ok acum?**. 8. La fel ca în exemplul precedent, Fig. 15, 16, 17, 18 reprezintă grafic componente imaginii din Fig. 14. Un adversar constată cu ușurință prin obținerea informațiilor stocate aflate pe un singur nod (indiferent de mărimea pragului k) că fișierul original conține un pattern care se repetă.

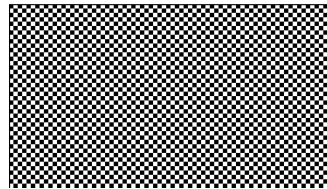


Fig. 14: Imagine conținând pattern-uri

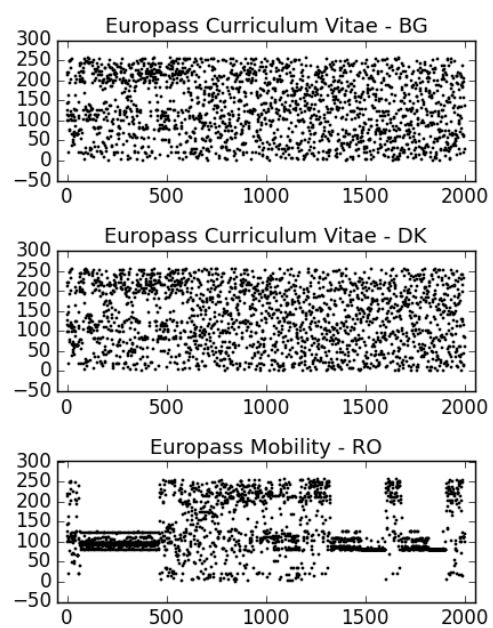


Fig. 12: Nod 3: Graficul componentelor pentru primele 2000 de blocuri

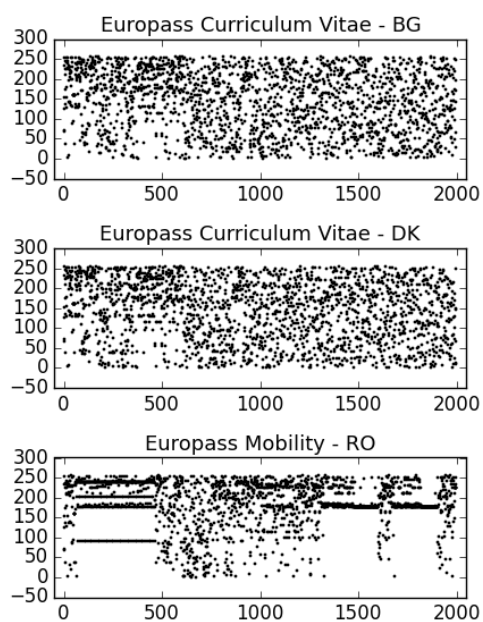


Fig. 13: Nod 4: Graficul componentelor pentru primele 2000 de blocuri

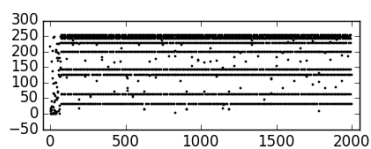


Fig. 15: Nod 1: Graficul componentelor pentru primele 2000 de blocuri

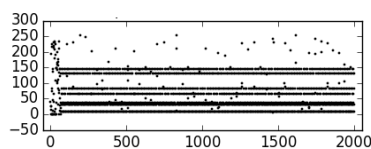


Fig. 16: Nod 2: Graficul componentelor pentru primele 2000 de blocuri

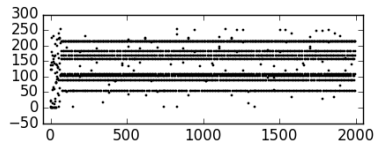


Fig. 17: Nod 3: Graficul componentelor pentru primele 2000 de blocuri

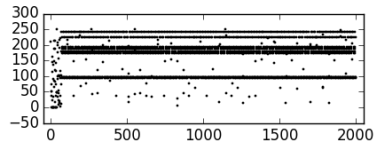


Fig. 18: Nod 4: Graficul componentelor pentru primele 2000 de blocuri

6.3 Publicarea rezultatelor

Pe baza rezultatelor obținute am redactat (împreună cu mentorul) un articol, aflat momentan în procesul de recenzie la Journal of Control Engineering and Applied Informatics (jurnal indexat ISI, categoria C). [28]

TODO Bibliografie:

1. Accesat ultima data - trebuie inlocuit cu Ultima accesare
 2. diffie hellman cu majuscule - in fisierul bib pui in { }ce vrei sa apara cu litere mari
 3. ai inca la bibliografie quantum - adauga info in text sau sterge de la bibliografie
- nu cred ca e utilizata

Referințe

1. WebSite: The Caesar cipher (2015) Ultima accesare: Februarie, 2015, <http://www.cs.trincoll.edu/crypto/historical/caesar.html>.
2. WebSite: Hacking the Vigenere cipher (2015) Ultima accesare: Februarie, 2015, <http://inventwithpython.com/hacking/chapter21.html>.
3. Katz, J., Lindell, Y.: Introduction to Modern Cryptography (Chapman & Hall/Crc Cryptography and Network Security Series). Chapman & Hall/CRC (2007)
4. Beimel, A.: Secret-sharing schemes: a survey. In: Coding and cryptology. Springer (2011) 11–46
5. Hillery, M., Bužek, V., Berthiaume, A.: Quantum secret sharing. *Physical Review A* **59**(3) (1999) 1829
6. Csirmaz, L.: The size of a share must be large. *Journal of cryptology* **10.4** (1997) 223–231
7. Martin, K.M.: Challenging the adversary model in secret sharing schemes. *Coding and Cryptography II, Proceedings of the Royal Flemish Academy of Belgium for Science and the Arts* (2008) 45–63
8. Boneh, D.: The decision Diffie-Hellman problem. In: *Algorithmic number theory*. Springer (1998) 48–63
9. Blakley, G.: Safeguarding cryptographic keys. *Proceedings of the 1979 AFIPS National Computer Conference* (1979) 313–317
10. Shamir, A.: How to share a secret. *Commun. ACM* **22**(11) (1979) 612–613
11. Ito, M., Saito, A., Nishizeki, T.: Secret sharing scheme realizing general access structure. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)* **72**(9) (1989) 56–64
12. Benaloh, J., Leichter, J.: Generalized secret sharing and monotone functions. In: *Proceedings on Advances in Cryptology. CRYPTO '88*, New York, NY, USA, Springer-Verlag New York, Inc. (1990) 27–35
13. Karnin, E.D., Member, S., Greene, J.W., Member, S., Hellman, M.E.: On secret sharing systems. *IEEE Transactions on Information Theory* **29** (1983) 35–41
14. Storer, M.W., Greenan, K.M., Miller, E.L., Voruganti, K.: Potshards - a secure, recoverable, long-term archival storage system. *TOS* **5**(2) (2009)
15. Shor, P.W.: Algorithms for quantum computation: discrete logarithms and factoring. In: *Foundations of Computer Science, 1994 Proceedings., 35th Annual Symposium on, IEEE* (1994) 124–134
16. Wylie, J.J., Bigrigg, M.W., Strunk, J.D., Ganger, G.R., Kiliççöte, H., Khosla, P.K.: Survivable information storage systems. *Computer* **33**(8) (2000) 61–68
17. Subbiah, A., Blough, D.M.: An approach for fault tolerant and secure data storage in collaborative work environments. In: *StorageSS*. (2005) 84–93
18. Patterson, D.A., Gibson, G., Katz, R.H.: A case for redundant arrays of inexpensive disks (raid). *SIGMOD Rec.* **17**(3) (June 1988) 109–116
19. Feldman, P.: A practical scheme for non-interactive verifiable secret sharing. In: *Proceedings of the 28th Annual Symposium on Foundations of Computer Science. SFCS '87*, Washington, DC, USA, IEEE Computer Society (1987) 427–438
20. Schwarz, T.J.E., Miller, E.L.: Store, forget, and check: Using algebraic signatures to check remotely administered storage. In: *26th IEEE International Conference on Distributed Computing Systems (ICDCS 2006)*, 4-7 July 2006, Lisboa, Portugal. (2006) 12
21. Alouneh, S., Abed, S., Mohd, B.J., Kharbutli, M.: An efficient backup technique for database systems based on threshold sharing. *JCP* **8**(11) (2013) 2980–2989

22. Vaudenay, S.: Security flaws induced by cbc padding - applications to ssl, ipsec, wtls. In: Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques: Advances in Cryptology. EUROCRYPT '02, London, UK, UK, Springer-Verlag (2002) 534–546
23. WebSite: Python Programming Language - official website (2015) Ultima accesare: Februarie, 2015, <https://www.python.org/>.
24. Hunter, J.D.: Matplotlib: A 2D graphics environment. Computing In Science & Engineering **9**(3) (2007) 90–95
25. WebSite: Cerealizer package (2015) Ultima accesare: Februarie, 2015, <https://pypi.python.org/pypi/Cerealizer>.
26. WebSite: Github - Official Website (2015) Ultima accesare: Februarie, 2015, <https://www.github.com>.
27. WebSite: Cod Github - Official Website (2015) Ultima accesare: Februarie, 2015, <https://www.github.com/rdragos/splitting-scheme>.
28. WebSite: Journal of Control Engineering and Applied Informatics (2015) Ultima accesare: Februarie, 2015, <http://www.ceai.srait.ro/index.php/ceai/index>.