2nd. International Conference on
Information Technology, Communications and Telecommunications (irICT2016)
1-2 March 2016 - Iran, Tehran

The two private information retrieval protocols and comparing them

In cryptography, Private Information Retrieval (PIR) is a protocol that allows users to retrieve data from the server that holds the database while the user's choice of data is not distinguishable.
In this paper, considering the importance of the privacy of users and servers, two protocols PIR are studied. PIR information theory (itPIR) that guarantee the privacy of user information theory and computational PIR (CPIR) guaranteed computing privacy. Furthermore, the relationship between them will be discussed as well as related problems and weaknesses. Moreover, the examples are presented in order to clarify the two protocols status.

Keywords: private information retrieval, privacy, PIR computing, PIR information theory.

بررسی دو پروتکل بازیابی خصوصی اطلاعات و مقایسه آنها

[۱]حوریه شاه‌حسینی ، [۲] سیدمحسن سائسی

[۱] کارشناسی ارشد دانشگاه تربیت مدرس، h.shahhoseini@gmail.com
[۲] کارشناسی ارشد دانشگاه تربیت مدرس، saessi.pub@gmail.com

چکیده :

در علم رمزنگاری، بازیابی خصوصی اطلاعات (PIR) پروتکلی می‌باشد که به کاربران امکان بازیابی داده‌ها را از سروری می‌دهد که پایگاه داده را در اختیار دارد و این درحالیست که داده انتخابی کاربر در حال بازیابی مشخص نیست. با توجه به اهمیت حریم خصوصی کاربران و سرورها در این مقاله دو پروتکل PIR اشاره شده که PIR تئوری اطلاعاتی (itPIR)، حریم خصوصی تئوری اطلاعاتی را به کاربر ضمانت می‌دهد و PIR محاسباتی (CPIR) حریم خصوصی محاسباتی را به کاربر تضمین می‌دهد و ارتباط بین آنها و همچنین مشکلات و نقاط ضعف آنها بررسی می‌شود. همچنین مثال‌هایی در این رابطه آورده شده است که به بیان بهتر این دو پروتکل می‌پردازد.

واژگان کلیدی : بازیابی خصوصی اطلاعات، حریم خصوصی، PIR محاسباتی، PIR تئوری اطلاعاتی.

۱- مقدمه

رشد اینترنت موجب فرصت‌های بسیاری برای محاسبات مشترک شد که افراد محاسبات را مبتنی بر ورودی‌های خصوصی خود مشترکاً و با هم هدایت می‌کنند. این محاسبات بین طرفین نامطمئن یا بین رقیبان می‌باشد. به عنوان نمونه مشتری‌ها ممکن است پرس‌وجوهایی به پایگاه داده‌های دوردست بفرستند که شامل اطلاعات خصوصی باشد و دو سازمان مالی رقیب با هم در یک پروژه سرمایه‌گذاری کنند که شامل محدودیت‌های با ارزش و خصوصی دو سازمان باشند. [1]

گاهی لازم است که به سیستم‌های محاسباتی به گونه‌ای ارجاع داده شود که چند طرف ارزش یکسانی را بر اساس بیت‌های رمزی اطلاعات منحصر به خود با هم محاسبه کنند، ولی رمزهای خود را برای طرفین در طی فرآیند آشکار نمی‌شود. برای مثال دو فردی که هر یک مالک اطلاعات رمزی x و y هستند ممکن است بخواهند به طور مشترک تابع f(x,y) را محاسبه کنند بدون آشکارسازی اطلاعاتی در مورد x و y غیر از آنچه که به طور معقول با دانستن مقدار واقعی f(x,y) استنباط شود. در محاسبات چند طرفه امن این مسئله در زمینه‌های مختلف بررسی می‌شود.

در مواردی که اهداف کاربر نگهداری یک راز باشد، اغلب در رابطه با دسترسی به پایگاه داده‌های عمومی هوشیار هستند. مثلاً پرس‌وجوی سرمایه‌گذاران یک پایگاه داده‌های بورس سهام و ارز برای ارزش بازار جاری بورس‌های خاص ممکن است میل به آشکارسازی علاقمندی‌هایشان در بورس نباشند، زیرا امکان دارد قیمت‌های آنها را تحت تأثیر قرار دهد. طرح‌های PIR پروتکل‌های رمزنگاری طراحی شده برای حمایت از حریم خصوصی کاربران پایگاه داده‌ها است. آنها به کلاینت‌ها امکان بازیابی رکوردها از پایگاه داده‌های عمومی را می‌دهند که هویت رکوردهای بازیابی شده از مالکان پایگاه داده‌ها پنهان‌سازی می‌شود. در این مقاله دو نوع

پروتکل‌های PIR را بررسی و به تفاوت و نقاط ضعف هر یک می‌پردازیم.

۲- کارهای مرتبط

در گذشته تحقیق در مورد محاسبات چند طرفه امن اغلب روی مطالعات تئوریکی متمرکز بود و مسائل کاربردی بسیاری مطالعه شده‌اند. نمونه‌های بسیاری از مسائل محاسبات چند طرفه امن نظیر مسأله بازیابی اطلاعات امن، پایگاه داده‌های آماری با حفظ حریم خصوصی و داده‌کاوی با حفظ حریم خصوصی در مورد آن وجود دارد. [4]

Chor و همکارانش طرح‌های PIR مطرح کردند که امکان بازیابی رکوردهای خصوصی را از پایگاه داده‌های تکراری با میزان کمی ارتباطات مهم بود. در این پروتکل‌ها، کاربران با هر سرور نگهدارنده پایگاه داده‌ها پرس‌وجو انجام می‌دهند. پروتکل مطمئن است که هر سرور با مشاهده تنها پرس‌وجویی که دریافت می‌کند هیچ اطلاعاتی در مورد هویت آیتم‌های کاربر ندارد [3].

طرح‌ها PIR (بازیابی خصوصی اطلاعات) توسط Chor ، Goldreich ، Kushilevitz و Sudan معرفی شد و امکان بازیابی اطلاعات از پایگاه داده‌ها بون افشاء علاقمندی‌های او به وجود آمد. ارزش این پروتکل‌ها به پیچیدگی ارتباطی آنهاست که برحسب تعداد بیت‌های انتقال یافته بین کاربر و سرورها تعیین می‌شود. [5]

پروتکل‌های PIR به دو دسته اصلی تقسیم می‌شوند که این تقسیم‌بندی بر حسب نوع حریم خصوصی است که به کاربر ضمانت داده می‌شود :

PIR تئوری اطلاعاتی (itPIR) : حریم خصوصی تئوری اطلاعاتی را به کاربر ضمانت می‌دهد نظیر حریم خصوصی سرورهای بدون محدودیت از نظر محاسباتی.

PIR محاسباتی (CPIR) : حریم خصوصی محاسباتی را به کاربر تضمین می‌دهد. نظیر حریم خصوصی علیه سرورهای محدود از نظر محاسباتی که نیاز به تعیین و تعریف فرضیات قابل تعامل مناسب داریم.

طرح‌های PIR تئوری اطلاعات و محاسبات متکی به مجموعه‌های مختلف از تکنیک‌ها هستند. برای بازیابی رکوردی از یک پایگاه داده، کاربر از دو سرور پرس‌وجو می‌کند. هر کدام یک کپی از پایگاه داده‌ها را ذخیره کرده‌اند، ولی پرس‌وجوهای منحصر بفرد، هیچ اطلاعاتی در مورد بیش از آنچه کاربر در جستجوی آن هست نگه نمی‌دارد. رکورد مورد نظر از پاسخ‌های ترکیبی سرور به دست می‌آید. در اینجا به شرح هر یک و پروتکل‌هایش می‌پردازیم. [3].

۳- PIR تئوری اطلاعاتی

فرض کنید که ۴ کپی از پایگاه داده‌ها وجود دارد و بیت‌های x به صورت ماتریس $n^{1/2} \times n^{1/2}$ مرتب شده‌اند. آلیس می‌خواهد X_{ij} را که $i,j<=n^{1/2}>=1$ است بازیابی کند.

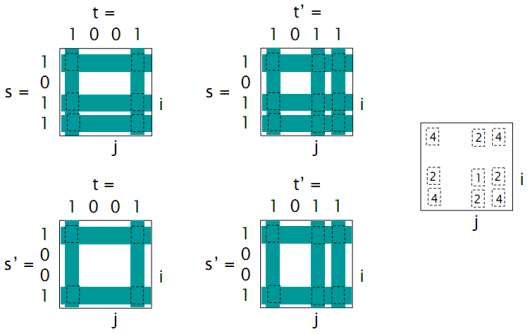
۳-۱ پروتکل

آلیس دو رشته بیت تصادفی s و t به طول $n^{1/2}$ تولید می‌کند. فرض 's همان s ای باشد که i امین بیت آن چرخیده (تبدیل ۰ با ۱ و بالعکس) و 't نیز همان t باشد که j امین بیت آن چرخیده است.

- آلیس موارد زیر را ارسال می‌کند :
  - s و t را به DB_0
  - s و t ' را به DB_1
  - s' و t را به DB_2
  - S' و t را به DB_3

هر DB یک بیت منفرد را به صورت XOR از بیت‌های X_{ab} محاسبه می‌کند که a امین بیت s (یا s') و b امین بیت t (یا t')برابر با ۱ است. آلیس بیت‌های دریافتی را XOR نموه و X_{ij} نتیجه می‌دهد.

۳-۲ نحوه عملکرد پروتکل PIR نمونه:



شکل ۱ چهار ماتریس و نتیجه نهایی پس از اعمال XOR

در اینجا هر پایگاه داده ۲ بردار تصادفی که مستقل از i و j است دریافت می‌کند. از طرفی هیچ اطلاعاتی در مورد i و j به پایگاه داده‌ها نمی‌رسد.

۳-۳ مثال عملی پروتکل PIR تئوری اطلاعاتی

فرض رشته بیت نمونه به صورت زیر باشد :

X=1010001100111100

در اینجا :

سطرهای ماتریس ۱۰۱۰ و ۰۰۱۱ و ۰۰۱۱ و ۱۱۰۰ هستند.

n=16 , xij=x22

یعنی x22 مورد نظر برای بازیابی است.

دو رشته تصادفی به ترتیب s و t و همینطور s' و t ' را تعریف می‌کنیم : برای s' و t ' به ترتیب بیت iام و jام چرخیده می‌شود

یعنی ۰ به ۱ و بالعکس)

S=1100 , t=0110
S'=1000 , t'=0010

اکنون :

S,t	1100,0110	x12 x13∂x22∂x23=0
S',t	1000,0110	x12∂x13=1
S,t'	1100,0010	x13∂x23=0
S',t'	1000,0010	x13=1

0∂1∂0∂1=0

پس بیتی که کاربر به دنبال آن است مقدار ۰ دارد و مشاهده می کنیم که در ماتریس همان می باشد.

۴- PIR محاسباتی

پروتکل‌های PIR تئوری اطلاعاتی هیچ اطلاعاتی (در حالت اطلاعات تئوری) در مورد اندیس درخواستی توسط آلیس را نشر نمی‌دهد. آنها در مقابل حمله‌ها حتی از یک پایگاه داده‌ها با توان محاسباتی نامحدود مقاومت می‌کنند. پروتکل‌های PIR محاسباتی (CPIR) ضمانت‌های ضعیف‌تری ارائه می‌کنند : آنها اطمینان می‌دهند که پایگاه داده‌ها هیچ اطلاعاتی به دست نمی‌آورد مگر این که مسأله سخت محاسباتی را حل کند (کاهش).

پروتکل‌های PIR تئوری اطلاعات به بیش از یک کپی غیر مرتبط از پایگاه داده‌ها نیاز دارند در حالی که پروتکل‌های CPIR با سربار ارتباطی کمتر حتی برای پایگاه داده‌های منفرد وجود دارد.

طرح‌های PIR محاسباتی شبیه نسخه دوم (تئوری اطلاعاتی آنها هستند، ولی ضمانت ضعیف‌تری ارائه می‌کند. آنها به طور خاص فقط اطمینان می‌دهند که یک سرور نمی‌تواند هیچ اطلاعات خاصی را در مورد گرایش کاربر بگیرد مگر این که مسأله سخت محاسباتی قطعی را حل کند (نظیر تجزیه عدد تصادفی بزرگ به عوامل اول). ارائه ضمانت‌های حریم خصوصی یا امنیت مبتنی بر فرض سختی محاسباتی، در رمزنگاری مدرن متداول است. در مقایسه با طرح‌های PIR تئوری اطلاعاتی، پروتکل‌های PIR محاسباتی (تحت فرضیات استاندارد) حتی وقتی پایگاه داده در یک سرور منفرد ذخیره شده است، سربار وجود دارد. با این وجود پارامترهای دنیای واقعی نمونه که پروتکل‌های محاسباتی معروف هستند کارآیی کمتری نسبت به اطلاعات تئوری شناخته شده دارند. مهم‌ترین مسئله در PIR محاسباتی اینست که تکرار داده‌ها برای مسأله PIR محاسباتی ضروری نیست. در اینجا مبتنی بر فرض باقیمانده مربعی، پایگاه داده‌های منفرد است. طرح‌های PIR امکان بازیابی اطلاعات از یک پایگاه داده‌ها را با حفظ حریم خصوصی پرس‌وجوها ارائه می‌دهند. به طور رسمی داده‌ها را به صورت یک رشته x n بیتی نمایش می‌دهیم که کاربر می‌خواهد بیت x_i را با حفظ شاخص i خصوصی از پایگاه داده‌ها به دست آورد.

۴-۱ سختی محاسبات در تئوری اطلاعاتی

مسأله زیر را در مورد رمزنگاری تئوری اطلاعاتی مجدداً بررسی می‌کنیم :

آیا پیچیدگی ارتباطی محاسبات امن غیر شرطی بستگی به پیچیدگی محاسباتی تابعی دارد که محاسبه می‌شود؟

برای مثال آیا اجراگرهایی که از نظر محاسباتی نامحدودند می‌توانند تابع اختیاری از ورودی‌های خود را با پیچیدگی ارتباطی چند جمله‌ای و آستانه خطی از حریم خصوصی غیر شرطی محاسبه کنند؟ آیا این کار را می‌توان با استفاده از تعداد ثابتی از ارتباطات انجام داد؟

طرح‌های PIR محاسباتی بسیار جذاب هستند زیرا آنها نیاز به نگهداری کپی‌های تکراری از پایگاه داده‌ها ندارند و با حریم خصوصی کار علیه تبانی سرورها همکاری نمی‌کنند. [8]

در اینجا بر "فرض سختی استاندارد فرض باقیمانده مربع" دلالت می‌کند. در عبارت $x2=a \pmod m$ فرض کنید m یک عدد صحیح باشد و عدد a باقیمانده مربع یا QR به پیمانه m باشد: در غیر این صورت a یک غیر باقیمانده مربع یا QNR به پیمانه m نامیده می‌شود.

فرض کنید QR نشان می‌دهد که محاسبات سخت اعدادی را که QR به پیمانه m هستند از آنهایی که این وِیژگی را ندارند تشخیص می‌دهند. مگر این که شخصی عوامل تجزیه m را بداند.

۴-۲ پروتکل

فرض کنید m یک عدد صحیح مثبت است. همچنین عدد a یک باقیمانده مربعی (QR) به پیمانه m است و

اگر $x^2 \equiv a \mod m$ $x \in Z$ و در غیر این صورت a یک باقیمانده غیر مربع (QNR) به پیمانه m است.

از نظر محاسباتی تشخیص اعداد QR از QNR مشکل است مگر این که تجزیه به عوامل اول m را بدانیم. در اینجا بیت‌های x به صورت ماتریس‌های $n^{1/2} \times n^{1/2}$ مرتب شده‌اند. آلیس می‌خواهد Xij $(j<=n^{1/2}, i>=1)$ را بازیابی کند.

۴-۳ پروتکل CPIR (همراه با عوامل تجزیه‌اش)

۴-۴ پروتکل

آلیس یک عدد تصادفی بزرگ m را به طور تصادفی انتخاب می‌کند. او $n^{1/2}$-1 عدد QR تصادفی تولید و QR های به پیمانه m را ایجاد می‌کند.

QNR mod $m=a_1, a_2, \ldots, a_{i-1}, a_{i+1}, \ldots$

سپس یک عدد QNR تصادفی mod $m=b_i$ تولید می‌کند و

$a_1, a_2, \ldots, a_{i-1}, b_i, a_{i+1}$,... را به پایگاه داده‌ها ارسال می‌کند.

در اینجا سرور نمی‌تواند تفاوت بین QNR ها و QR ها به پیمانه m را تشخیص دهد بنابراین از نقطه دید سرور، بردار دریافتی فقط آرایه‌ای از اعداد تصادفی u_1, u_2, ... است.

در نهایت او برای هر ستون c از x ، پایگاه داده‌ها عبارت $V_c = u_1^{x1c} u_2^{x2c} \ldots \bmod m$ را محاسبه می‌کند :

پایگاه داده‌ها با v_1, v_2,... پاسخ می‌دهد. آلیس بررسی می‌کند که آیا V_j یک QR یا یک QNR به پیمانه m است.

اگر QR بود $x_{ij}=0$ و اگر QNR بود $x_{ij}=1$

۵-۴ نحوه عملکرد پروتکل CPIR :

$$X = \begin{matrix} \cdots & x_{1j} & \cdots \\ \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{ij} \cdots \\ \cdots & & \end{matrix} \qquad U = \begin{matrix} a_1 \\ \cdots \\ b_i \\ \cdots \end{matrix} \qquad v_j = a_1^{x_{1j}} \ldots b_i^{x_{ij}} \ldots \bmod m$$

شکل ۲

اگر xij=0 فقط QR ها ضرب می‌شوند در غیر این صورت QR ها با یک QNR منفرد ضرب می‌شوند.

مشخص است که QR×QNR=QNR و QR×QR=QR

۶-۴ مثال عملی برای پروتکل PIR محاسباتی

فرض رشته بیت به صورت زیر باشد :

X=1010001100111100

در اینجا سطرهای ماتریس ۱۰۱۰ و ۰۰۱۱ و ۱۱۰۰ هستند.

ابعاد ماتریس نمونه ۴ در ۴ می‌شود.

n=16 , xij=x22

یعنی x22 مورد نظر برای بازیابی است.

حال باید یک m در نظر بگیریم. فرض m=15.

اعداد QR آن به ترتیب ۱و ۴ و ۶ و ۹ و ۱۰ هستند که باید bi را در همین محدود QNR در نظر گرفت.

a1	b2	a3	a4	a5	a6
1	2	4	6	9	10
u1	u2	u3	u4	u5	u6

c=1 $v1=u_1^{x11}*u_2^{x21}*u_3^{x31}*u4^{x41}=1^{1*}2^{0*}4^{0*}6^{1}=6$

c=2 $v2= u_1^{x12}*u_2^{x22}*u_3^{x32}*u4^{x42}=1^{0*}2^{0*}4^{0*}6^{1}=6$

c=3 $v3= u_1^{x13}*u_2^{x23}*u_3^{x33}*u4^{x43}=1^{1*}2^{1*}4^{1*}6^{0}=8$

c=4 $v4= u_1^{x14}*u_2^{x24}*u_3^{x34}*u4^{x44}=1^{0*}2^{1*}4^{1*}6^{0}=1$

برای X_{22} مقدار V_2 را بررسی می کنیم که در اینجا ۶ است و QR می‌باشد پس مقدار X_{22} برابر با صفر است.

۵- نتیجه‌گیری

از آنجا که شرح داده شد، PIRپروتکلی می‌باشد که به کاربران اجازه می‌دهد تا یک داده را از سروری که پایگاه داده را در اختیار دارد بازیابی کرده بدون این که معلوم شود که کدام داده که کاربر انتخاب کرده‌است در حال بازیابی است. طرح‌های PIR پروتکل‌های رمزنگاری طراحی شده برای حمایت از حریم خصوصی کاربران پایگاه داده‌هاست. آنها به کلاینت‌ها امکان بازیابی رکوردها از پایگاه داده‌های عمومی را می‌دهند که هویت رکوردهای بازیابی شده از مالکان پایگاه داده‌ها پنهان‌سازی می‌شود در اینجا ریسک مهمی برای حریم خصوصی کاربر وجود دارد زیرا مالک پایگاه داده‌های آلوده ممکن است پرس‌وجوهای کاربر را بررسی کرده و آنچه را که کاربر از آن پس سروکار دارد آلوده کند.

پروتکل‌های PIR به دو دسته اصلی تقسیم می‌شوند که تقسیم‌بندی آن بر حسب نوع حریم خصوصی است که به کاربر ضمانت داده می‌شود :

PIR تئوری اطلاعاتی (itPIR) : حریم خصوصی تئوری اطلاعاتی را به کاربر ضمانت می‌دهد نظیر حریم خصوصی سرورهای بدون محدودیت از نظر محاسباتی.

PIR محاسباتی (CPIR) : حریم خصوصی محاسباتی را به کاربر تضمین می‌دهد نظیر حریم خصوصی علیه سرورهای محدود از نظر محاسباتی که در این حالت ما نیاز به تعیین و تعریف فرضیات قابل تعامل مناسب داریم.

تفاوت اصلی بین این دو نوع اینست که حریم خصوصی تئوری اطلاعاتی به طور مؤثر تنها در صورتی امکان‌پذیر است که فقط پایگاه داده‌ها در k≥2 سرور غیر مرتبط تکرار شوند، اگر فقط یک سرور منفرد پایگاه داده‌ها را نگه دارد بهترین راه حل itPIR ، یک راه حل بدیهی است. مسأله PIR فقط مرتبط با حریم خصوصی کاربر است بدون نیاز به حفاظت از حریم خصوصی سرورها، پروتکل‌های PIR می‌توانند امکان کسب xi و برخی اطلاعات اضافی دیگر نظیر سایر بیت‌های پایگاه داده‌ها یا XOR بیت‌هایی از x را به کاربر بدهند. [6]

در نهایت، تفاوت‌های دو رویکرد از PIR را بررسی کردیم که PIR تئوری اطلاعاتی {CGKS95,Amb97,...} در آن پایگاه داده‌ها را بین k سرور تکرار می‌کند و اشکال آن در این بود که سرورها تبانی می‌کنند. PIR محاسباتی (CG97, K097, CMS99,...) در آن حریم خصوصی محاسباتی مبتنی بر فرضیات رمزنگاری- مسئله NP برای شکستن روش سخت بود.

هسته محیط‌های محاسباتی شبکه‌بندی شده (اینترنت، اینترانت)که اطلاعات را عبور می‌دهند در وضعیتی هستند که چندین سرورهای اطلاعاتی را به صورت عمومی توزیع می‌کنند.

همانطور که بررسی کردیم تلاش‌های بسیاری برای یافتن روش

محافظت از حریم خصوصی سرورها وجود دارد، مثلاً حفاظت سرورها از کاربران غیر قانونی (با تصدیق اصالت کاربران) یا از استراق‌سمع کننده‌ها (مثلاً با رمزنگاری) وجود داشت، ولی موضوع حفاظت از حریم خصوصی کاربران در مقابل سرورها، کاملاً جدید و نوظهور است.

۶- مراجع

[1] Dong C. and Chen L. , *A Fast Single Server Private Information Retrieval Protocol with Low Communication Cost* , ESORICS 2014 , Lecture Notes in Computer Science. Vol. 8712, pp 380-399 , 2014

[2] Fazeli A. , Vardy A., Yaakobi E. , *PIR with Low Storage Overhead: Coding instead of Replication* , 2015 , arXiv.org , cs , arXiv:1505.06241

[3] Feng Zhang L. and Safavi-Naini R., *Verifiable Multi-server Private Information Retrieval , Applied Cryptography and Network Security* , Lecture Notes in Computer Science Vol. 8479, pp 62-79 , 2014

[4] Kumar Mishra D. et al , *A Glance at Secure Multiparty Computation for Privacy Preserving Data Mining* , International Journal on Computer Science and Engineering Vol.1 , No 3, pp. 171-175, 2009

[5] Mikhail Atallah J. and Wenliang Du , *Secure MultiParty Computation Problems and Their Applications: A Review and Open Problems* , In Proceedings of the 2001 on New Security Paradigms , pp 11-20, New Mexico, USA, 2001.

[6] Wang J., Cellary W., Wang D., etc , Web Information Systems Engineering , WISE 2015, Springer,pp 509-511, 2015

[7] Yekhanin S. , *Private Information Retrieval, Cryptographic protocols safeguard the privacy of user queries to public* databases , Communications of the ACM, Vol. 53 No. 4, pp 68-73 , 2010

[8] Yekhanin S., *Private Information Retrieval*, Communications of the ACM, Vol. 53 No. 4, April 2010

[۹]خزائی، ش. ، مقدمه پیشرفته بر رمزنگاری ، دانشگاه صنعتی شریف ۱۳۹۱

[۱۰]دلدار، ف. ، رمزنگاری همریختی و کاربرد آن در بازیابی خصوصی اطلاعات رمزنگاری همریختی ، ۱۳۹۳