چهارمین کنفرانس ملی ایده‌های نو در مهندسی برق
۲۰و۲۱ آبان ماه ۱۳۹۴ - دانشگاه آزاد اسلامی واحد اصفهان (خوراسگان)

حفظ حریم خصوصی یال در خوشه‌بندی داده‌های منتشر شده‌ی شبکه‌های اجتماعی

داریوش عسگری[۱] و محمدرضا ابراهیمی دیشابی[۲]

[۱]گروه کامپیوتر، واحد زنجان، دانشگاه آزاد اسلامی، زنجان، ایران، dariush2914@gmail.com

[۲]گروه کامپیوتر، واحد میانه، دانشگاه آزاد اسلامی، میانه، ایران، mrebrahimy@m-iau.ac.ir

چکیده - در سالهای اخیر، شبکه‌های اجتماعی رشد قابل توجهی داشته و داده‌های زیادی در آنها تولید می‌شوند که حاوی اطلاعات مفیدی برای تحلیلگران پدیده‌های مختلف اجتماعی هستند. امنیت و حفظ حریم خصوصی داده‌های شبکه‌های اجتماعی، امری ضروری است؛ به خصوص زمانی که داده‌های یک شبکه‌ی اجتماعی، به‌منظور تحقیقات، بطور کامل منتشر می‌شوند. الگوریتم‌های مختلفی برای حفظ حریم خصوصی داده‌های منتشر شده‌ی شبکه‌های اجتماعی وجود دارند. دراکثر این الگوریتم‌ها، حریم خصوصی داده‌های منتشر شده با استفاده از دانش پیش‌زمینه کاربران، نقض می‌شود. به منظور رفع این مشکل، مفهوم حریم خصوصی تفاضلی ابداع شد که مستقل از دانش پیش زمینه کاربر تعریف می‌شود. در این مقاله، برای حفظ حریم خصوصی یال در خوشه‌بندی داده‌های منتشر شده‌ی شبکه‌های اجتماعی، الگوریتمی مبتنی بر مفهوم حریم خصوصی تفاضلی ارائه شده است. در الگوریتم ارائه شده، به منظور افزایش کارایی داده‌های منتشر شده، از تبدیل موجک گسسته هار استفاده می‌شود. در این مقاله، حریم خصوصی تفاضلی داده‌های منتشر شده را با استفاده از مدل ریاضی به اثبات خواهیم رساند. همچنین با آزمایش الگوریتم ارایه شده بر روی تعدادی از مجموعه داده‌های شناخته شده، نشان خواهیم داد که روش پیشنهادی، از درجه کارایی بالاتری نسبت به روشی که اخیرا ارایه شده است، برخوردار است.

کلید واژه- شبکه اجتماعی، حریم خصوصی تفاضلی، خوشه‌بندی، تبدیلات موجک گسسته

۱- مقدمه

گسترش استفاده از شبکه‌های اجتماعی، نوع جدیدی از ارتباطهای اجتماعی را با استفاده از فن‌آوری اطلاعات در فضای مجازی فراهم کرده است. تحلیگران برای تجزیه و تحلیل پدیده‌های مختلف اجتماعی مانند اپیدمیولوژی (Epidemiology)، بازاریابی و غیره از داده‌های تولید شده در شبکه‌های اجتماعی استفاده می‌کنند [1]. از طرفی، حفظ حریم خصوصی داده‌های منتشر شده شبکه‌های اجتماعی، یکی از مهمترین مسائلی است که توجه زیادی به آن می‌شود. اکثر الگوریتم‌های ارائه شده در این زمینه، در مقابل دانش پیش‌زمینه کاربران آسیب‌پذیر هستند. از این رو مفهوم حریم خصوصی تفاضلی ابداع شد [2] . هدف اصلی در حریم خصوصی تفاضلی این است که یک کاربر با هر مقدار دانش پیش‌زمینه خود، توانائی نقض حریم خصوصی داده‌ها را نداشته باشد. یکی از مهمترین ایرادهای حریم خصوصی تفاضلی، کارایی پائین آن به دلیل اضافه کردن خدشه (noise) به داده‌های نهایی است. بنابراین، باید به دنبال روشی باشیم تا علاوه بر حفظ حریم خصوصی داده‌ها، کارایی بالایی را هم برای داده‌های منتشر شده داشته باشیم. در این مقاله، الگوریتمی برای حفظ حریم خصوصی یال (edge) در خوشه‌بندی داده‌های منتشرشده شبکه اجتماعی ارائه شده است. به منظور افزایش کارایی داده‌های منتشر شده، از تبدیل موجک گسسته هار استفاده شده است. استفاده از تبدیل موجک هار باعث می‌شود تا (۱) خدشه کمتری را به منظور رسیدن به شرایط حریم خصوصی تفاضلی، به داده‌ها اضافه کنیم از این‌رو، داده‌های منتشر شده، کارایی بالایی را خواهند داشت و (۲) اندازه ابعاد داده‌های منتشر شده به مراتب کمتر از اندازه ابعاد داده‌های اولیه خواهد بود که باعث می‌شود تا الگوریتم‌های داده‌کاوی با سرعت و کارایی بالاتری بر روی داده‌های منتشر شده اجرا شوند. حریم خصوصی تفاضلی داده‌های منتشر شده را با مدل ریاضی اثبات خواهیم کرد. همچنین با آزمایش الگوریتم ارایه شده بر روی تعدادی از مجموعه داده‌های شناخته شده، نشان خواهیم داد که روش پیشنهادی، از درجه کارایی بالاتری نسبت به روشی که اخیرا ارایه شده است، برخوردار است.

۲- مفاهیم اولیه

۲-۱- تبدیل موجک هار:

در تبدیل موجک (غیر نرمال) هار، فیلترهای پایین‌گذر و بالاگذر به ترتیب عبارتند از $\{h_0, h_1\}$ و $\{g_0, g_1\}$ که در آن $h_0 = h_1 = \frac{1}{2}, g_0 = \frac{1}{2}, g_1 = -\frac{1}{2}$ هستند. مطابق با الگوریتم «مالات و مایر» (Mallat & Meyer) [3]، اگر فیلترهای پایین‌گذر (بالاگذر) را بر روی 2^l ضریب تقریب واقع در سطح l اعمال کنیم در این صورت، 2^{l-1} ضریب تقریب (موجک) در سطح $l-1$ را به دست خواهیم آورد. این الگوریتم به صورت بازگشتی تا سطح صفر اجرا شده و ضرایب تقریب و موجک را در سطوح پایین‌تر محاسبه خواهد کرد. در این تبدیل، اندازه داده باید توانی از ۲ باشد.

۲-۲- حریم خصوصی تفاضلی:

تعریف (۱) [4]: مکانیزم A خاصیت حریم خصوصی تفاضلی- (ϵ, δ) دارد هرگاه برای تمام مجموعه داده‌های T و Ť (که فقط در یک مشخصه با هم تفاوت دارند) و تمام مجموعه داده‌های خروجی $\hat{D} \subseteq \text{Range}(A)$، رابطه‌ی زیر برقرار باشد:

$$\Pr[A(T) \in \hat{D}] \leq e^{\epsilon}.\Pr[A(\acute{T}) \in \hat{D}] + \delta$$

در حریم خصوصی تفاضلی-ϵ، مقدار $\delta = 0$ است. حساسیت-L_1 (L1_Sensitivity) (یا حساسیت سراسری) یکی از کلیدی‌ترین مفاهیمی است که در ایجاد داده‌های تفاضلی از آن استفاده شده است. در [5]، حساسیت-L_1 مجموعه‌ای از توابع به صورت زیر تعریف شده است (با اعمال تغییر کوچک در آن):

تعریف (۲): فرض کنید F نشان دهنده‌ی مجموعه‌ای از توابع حقیقی f (یعنی خروجی آنها یک عدد حقیقی است) باشد. در این صورت حساسیت-L_1 مجموعه F عبارت است از کوچکترین عددی مانند $S(F)$ به طوری که شرط $\underset{T,\acute{T}}{\text{Max}} \sum_{f \in F} |f(T) - f(\acute{T})| \leq S(F)$ برای آن برقرار باشد که در آن T و $\acute{T}$ دو مجموعه داده‌ای هستند که فقط در یک مشخصه از یک رکورد با هم تفاوت دارند.

۳- کارهای مرتبط

مفهوم حریم خصوصی تفاضلی توسط DWORK ابداع شد [2].

در مرجع [4]، از رویکرد ماتریس تصادفی (random matrix) به منظور ایجاد داده‌های تفاضلی از شبکه‌های اجتماعی استفاده شده است. ابتدا داده‌های شبکه تبدیل به یک ماتریس مجاورت (Adjacency Matrix) در ابعاد $n \times n$ می‌شوند. سپس یک ماتریس تصادفی با ابعاد $n \times m$ که درایه‌های آن دارای توزیع تصادفی نرمال (normal random distribution) هستند را در ماتریس مجاورت ضرب گرده و در نهایت، خدشه‌هایی با توزیع نرمال را به درایه‌های ماتریس به دست آمده اضافه می‌نمایند. ایراد اصلی این رویکرد، استفاده از مقدار بالای خدشه است که سبب می‌شود تا داده‌های منتشر شده از سودمندی مناسبی برخوردار نباشد. در مرجع [6] از روش گراف طیفی (spectral graph) برای حفظ حریم خصوصی استفاده شده است. ابتدا گراف شبکه اجتماعی تبدیل به یک ماتریس مجاورت شده سپس بردارها و مقادیر ویژه ماتریس مجاورت محاسبه می‌شوند. در نهایت، خدشه‌هایی با توزیع لاپلاس (مکانیزم لاپلاس) را بر روی مقادیر ویژه و هر ورودی بردارهای ویژه اضافه می‌نمایند. در این روش، کیفیت خوشه‌بندی داده‌های منتشر شده ضعیف است. در مرجع [7]، پیشنهاد انتشار ماتریس کواریانس از داده‌های اصلی ترکیب شده با خدشه‌های تصادفی را ارائه کرده است. این رویکرد برای انتشار داده‌های شبکه، نیاز به محاسبات سنگین برای درایه‌های ماتریس شبکه‌های اجتماعی و همچنین نیاز به فضای ذخیره‌سازی انبوهی خواهد داشت. در مرجع [8]، از تبدیلات تبدیلات لیندن اشتراس (Johnson–Lindenstrauss) برای ایجاد داده‌هایی تفاضلی استفاده شده است. نیاز به محاسبات بسیار بالا و مصرف حافظه زیاد از مهمترین ایرادهای این روش است.

۴- الگوریتم پیشنهادی

در این بخش، الگوریتم ارایه شده را معرفی خواهیم کرد. به منظور راحتی، الگوریتم پیشنهاد شده را EDP (edge differential privacy) نامگذاری می‌کنیم. شکل ۱، شبه کد الگوریتم ارایه شده را نشان می‌دهد.

خط ۱، فایل متنی حاوی اطلاعات شبکه‌ی اجتماعی (M) را دریافت کرده و در خط ۲، آن را به ماتریس T با ابعاد $n \times n$ که فقط درایه‌های صفر یا یک دارد (۱ یعنی ارتباط بین دو گره

وجود دارد و صفر یعنی وجود ندارد) می‌کند. در خط ۳، ستون‌های ماتریس T، به توان ۲ رسانده می‌شود (بخش ۲-۱). برای این منظور، به تعداد $\acute{n} - n$ عدد صفر به ستون‌های ماتریس اولیه اضافه می‌شود تا ماتریس جدید $G^{n\times\acute{n}}$ با تعداد ستون‌های توانی از ۲ به دست آید که در آن، $\acute{n}$ نیز کوچکترین عدد بزرگتر یا مساوی n هست که توانی از ۲ می‌باشد. خطوط ۴، ۵ و ۶ تبدیلات موجک هار را روی سطر به سطر ماتریس G اعمال می‌نمایند تا به ماتریسی با ابعاد $n \times m$ برسیم. این حلقه، بر روی هر رکورد به صورت بازگشتی از سطح log($\acute{n}$) تا سطح log(m)+1 اجرا شده و در نهایت، ضرایب تقریب در سطح log(m) را به عنوان خروجی هر رکورد در ماتریس G_T با ابعاد $n \times m$ ذخیره می‌نماید (خط ۷). در خط ۸، ماتریسی تحت عنوان N در ابعاد $n \times m$ با درایه‌هایی از نوع خدشه‌های تصادفی لاپلاس با میانگین صفر و دامنه‌ی $\lambda = \frac{1}{\epsilon\times 2^{(\log(\acute{n})-\log(m))}}$ تشکیل می‌گردد. در خط ۹، ماتریس ضرائب تقریب (G_T) و ماتریس خدشه‌های تصادفی N با هم ترکیب می‌شوند. خط ۱۰، ماتریس نهایی ($\acute{G}$) برای استفاده در الگوریتم‌های داده‌کاوی منتشر می‌گردد.

Algorithm 1: EDP

Input: M: the social network information in Text format; m: the published dataset dimension which is power of 2

Output: $\acute{G} \in R^{n\times m}$: published dataset

1. **Function** [$\acute{G}$]= **EDP**(M, m)
2. Convert the input text M into $T \in R^{n\times n}$, a matrix representation.
3. Increase the length of each record $r \in T$ by adding $(\acute{n} - n)$ zero values to the end of record r, where $\acute{n}$ is the smallest power 2 value greater than or equal to n. Consider the resulting value is $G \in R^{n\times\acute{n}}$.
4. **foreach** record $r \in G$ **do**
5. **for** $l = \log(\acute{n})$ **downto** $'\log(m) + 1'$
6. Apply the Haar wavelet transform on the approximation coefficients of record r in the level 'l' to obtain the approximation coefficients of the level '$l - 1$'.
7. Add the resulting approximation coefficients to the matrix $G_T \in R^{n\times m}$.
8. Construct a $n \times m$ noise matrix N, where each element of N has been drawn from the Laplace distribution with mean zero and magnitude $\lambda = \frac{1}{\epsilon\times 2^{(\log(\acute{n})-\log(m))}}$
9. Construct $\acute{G} = G_T + N$.
10. Publish $\acute{G}$.

شکل ۱: الگوریتم EDP

۵- تحلیل حریم خصوصی الگوریتم ارائه شده

در ابتدا قضیه زیر را داریم:

قضیه (۱): [5] اگر F، مجموعه‌ی از توابع حقیقی با حساسیت $S(F)$ باشد (تعریف ۲) و اگر A الگوریتمی باشد که به خروجی هرکدام از این توابع، خدشه‌ی با توزیع لاپلاس که دارای میانگین صفر و دامنه‌ی λ هستند را اضافه کند در این صورت مکانیزم A خاصیت «حریم خصوصی تفاضلی- $\left(\frac{S(F)}{\lambda}\right)$» خواهد داشت.

اکنون قضیه زیر را ثابت می‌کنیم:

قضیه (۲): الگوریتم ارائه شده EDP (شکل ۱) خاصیت حریم خصوصی تفاضلی-ϵ دارد.

اثبات: فرض کنید $F = \{f_{11}, f_{12} \dots, f_{1m}, \dots, f_{nm}\}$ مجموعه توابعی باشد که ماتریس T را به ماتریس G_T، تبدیل می‌نماید. هر کدام از توابع داخل مجموعه F، را می‌توان تابعی در نظر گرفت که فقط یکی از درایه‌های رکورد حاصل از اجرای خطوط ۲ تا ۷ الگوریتم ارایه شده (شکل ۱) بر روی یک رکورد از ماتریس T را به خود اختصاص می‌دهند. در این صورت داریم:

$$T = \begin{pmatrix} r_{11} & r_{12} & \cdots & \cdots & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & \vdots & \vdots & r_{n2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ r_{i1} & r_{i2} & \cdots & r_{ij} & \cdots & r_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & \cdots & \cdots & r_{nn} \end{pmatrix} \xrightarrow{f_{11}\cdots f_{nm}} G_T = \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1m} \\ g_{21} & g_{22} & \cdots & g_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ g_{i1} & g_{i2} & \vdots & g_{im} \\ \vdots & \vdots & \vdots & \vdots \\ g_{n1} & g_{n2} & \cdots & g_{nm} \end{pmatrix}$$

که در آن:

$$f_{11}(T) = g_{11}, f_{12}(T) = g_{12}, \dots, f_{ij}(T) = g_{ij}, \dots, f_{nm}(T) = g_{nm}$$

حال یکی از درایه‌های ماتریس T را به اندازه‌ی (δ) تغییر داده و آن را $\acute{T}$ می‌نامیم. بدون اینکه خللی در اثبات قضیه پیش بیاید، فرض می‌کنیم این تغییر در درایه مربوط به سطر iام و ستون jام اتفاق افتاده باشد. با اعمال مجموعه توابع F برروی ماتریس $\acute{T}$ (اعمال خطوط ۲ تا ۷ شکل ۱)، ماتریس جدیدی تحت عنوان $\acute{G}_T$ به دست خواهد آمد.

$$\acute{T} = \begin{pmatrix} r_{11} & r_{12} & \cdots & \cdots & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & \vdots & \vdots & r_{n2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ r_{i1} & r_{i2} & \cdots & r_{ij} + \delta & \cdots & r_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & \cdots & \cdots & r_{nn} \end{pmatrix} \xrightarrow{f_{11}\cdots f_{nm}} \acute{G}_T$$

با توجه به اینکه، تبدیل موجک هار به طور مستقل بر روی هر

رکورد اعمال می‌شود در این صورت، رکود iام از ماتریس $\acute{G}_T$ فقط در یک درایه با رکورد iام از ماتریس G_T متفاوت خواهد بود و بقیه رکوردهای هر دو ماتریس، مقادیر یکسانی را خواهند داشت. بنابراین با توجه به مفهوم حساسیت سراسری، حساسیت مجموعه توابع F، برابر با مقدار $|\acute{G}_T - G_T|_1$ خواهد بود. از طرفی با توجه به اینکه این دو ماتریس فقط در یک درایه با هم فرق دارند بنابراین، تفاضل نظیر به نظیر درایه‌های دو ماتریس به غیر از درایه مربوط به سطر iام، برابر با صفر خواهد بود. به عبارت دیگر $|\acute{G}_T - G_T|_1 = |(g_{i1}, g_{i2}, \dots, g_{im}) - (\acute{g}_{i1}, \acute{g}_{i2}, \dots, \acute{g}_{im})|_1$ خواهد شد. برای محاسبه مقدار $|(g_{i1}, g_{i2}, \dots, g_{im}) - (\acute{g}_{i1}, \acute{g}_{i2}, \dots, \acute{g}_{im})|_1$ ، کافی است تا تبدیل موجک هار را تا سطح $\log(m) + 1$ (مطابق با کد ۵، شکل ۱) بر روی رکورد iام از ماتریس $\acute{T}$ (مطابق با خطوط ۲ تا ۷ الگوریتم ارایه شده) اعمال نماییم. به دلیل اینکه، رکورد iام ماتریس $\acute{T}$ فقط در یک درایه با رکورد iام ماتریس T متفاوت است از این‌رو، تفاوت نرم ۱ ($|.|_1$) ضرایب تقریب سطح kام مربوط به رکورد iام از ماتریس $\acute{T}$ نسبت به ضرایب تقریب سطح kام مربوط به رکورد iام از ماتریس T برابر با $\frac{\delta}{2^{(\log(\acute{n})-k)}}$ خواهد شد. مطابق با الگوریتم ۱ (کد ۷)، خروجی مورد نظر، ضرایب تقریب در سطح log(m) است. بنابراین تفاوت نرم ۱ ($|.|_1$) ضرایب تقریب سطح log(m)ام مربوط به رکورد iام از ماتریس $\acute{T}$ نسبت به ضرایب تقریب سطح log(m)ام مربوط به رکورد iام از ماتریس T برابر است با $\frac{\delta}{2^{(\log(\acute{n})-\log(m))}}$. به‌عبارت دیگر حساسیت سراسری الگوریتم پیشنهادی برابر با $\frac{\delta}{2^{(\log(\acute{n})-\log(m))}}$ است. اگر ماکزیمم مقدار $|\delta| = 1$ باشد آنگاه حساسیت سراسری برابر با $\frac{1}{2^{(\log(\acute{n})-\log(m))}}$ خواهد بود. پس $S(f) = \frac{1}{2^{(\log(\acute{n})-\log(m))}}$ خواهد شد.

حال با استفاده از قضیه ۱، اگر توزیع لاپلاس با دامنه‌ی λ را درنظر بگیریم به طوری‌که $\epsilon = \frac{S(f)}{\lambda}$ باشد آنگاه الگوریتم پیشنهادی خاصیت حریم خصوصی تفاضلی- (اپسیلون) را خواهد داشت. پس:

$$\epsilon = \frac{\frac{1}{2^{(\log(\acute{n})-\log(m))}}}{\frac{\lambda}{1}} = \frac{1}{\lambda.2^{(\log(\acute{n})-\log(m))}} \Longrightarrow \lambda = \frac{1}{\epsilon.2^{(\log(\acute{n})-\log(m))}}$$

۶- نتایج ارزیابی الگوریتم پیشنهادی

تنظیمات: برای ارزیابی الگوریتم ارائه شده از مجموعه داده‌های مربوط به کتابخانه SNAP [9] استفاده شده است (جدول ۱).

جدول (۱): مشخصات مجموعه داده‌های مورد آزمایش

مجموعه داده	تعداد راس	تعداد یال
Facebook- combined	۴۰۳۹	۸۸۲۳۴
P2p-Gnutella05	۸۸۴۶	۳۱۸۳۹
P2p-Gnutella08	۶۳۰۱	۲۰۷۷۷
Wiki-Vote	۷۱۱۵	۱۰۳۶۸۹

همچنین، الگوریتم پیشنهادی (EDP) را با الگوریتم The Random-Matrix algorithm (RMA) بر اساس معیارهای «حریم خصوصی» و «کیفیت خوشه‌بندی» مقایسه خواهیم کرد. شبه کد الگوریتم RMA در شکل ۲ بیان شده است.

Algorithm 4 (RMA): $\acute{A}$= Publish (A, m, σ^2)

Input : (1) symmetric adjacency matrix $A \in R^{n\times n}$
(2) the number of random projections $m < n$
(3) variance for random noise σ^2
Output : $\acute{A}$
1. Compute a random projection matrix P, with $p_{i,j} \sim N(0, 1/m)$
2. Compute a random perturbation matrix Q, with $Q_{i,j} \sim N(0, \sigma^2)$
3. Compute the projected matrix $A_p = A.P$
4. Compute the randomly perturbed matrix $\acute{A} = A_p + Q$

شکل ۲: الگوریتم RMA [4]

برای خوشه‌بندی داده‌های اولیه و داده‌های تولید شده توسط الگوریتم‌های EDP و RMA، از دو الگوریتم Spectral Clustering (SC) (شکل ۳) و Differential Private Spectral Clustering (DPSC) (شکل ۴) استفاده می‌کنیم. استفاده از الگوریتم SC برای داده‌های شبکه‌های اجتماعی بسیار مناسب است چرا که ورودی آن به جای نماینده داده، ماتریس مجاورت است.

Algorithm 2: Spectral Clustering

Input : (1) Adjacency Matrix $G \in R^{n\times n}$
(2) Number of clusters k
Output : clusters $C_1, \dots, C_k$
1. Compute first k eigenvectors $u_1, \dots, u_2$ of G
2. Get matrix $U \in R^{n\times k}$ where ith column of U is U_k
3. Obtain cluster by applying k-means clustering on matrix U

شکل ۳: الگوریتم SC [4]

Algorithm 3: Differential Private Spectral Clustering (DPSC)

Input: (1) Published matrix $G' \in R^{n\times m}$
(2) number of clusters k
Output: Clusters $C_1, \dots, C_k$

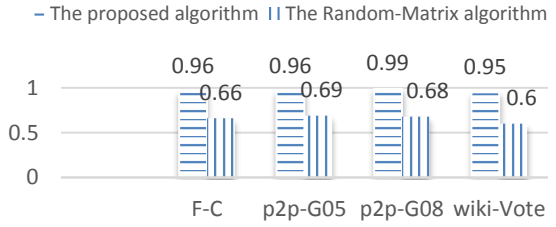
1. Compute first k eigenvectors $u_1, \dots, u_k$ of $\hat{G}$
2. Get matrix $U \in R^{n\times k}$ where ith column of U is U_k
3. Obtain cluster by applying k-means clustering on matrix U

شکل ۴: الگوریتم DPSC [4]

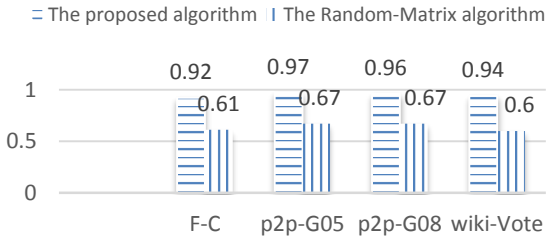
بعد از خوشه‌بندی داده‌های اولیه و داده‌های نهایی، با استفاده از معیارهای F-Measure [10] و NMI [11]، کیفیت خوشه‌بندی الگوریتم‌های EDP و RMA را محاسبه می‌کنیم. به دلیل اضافه شدن خدشه به داده‌های نهایی، این داده‌ها نیز تصادفی خواهند بود. از این‌رو، الگوریتم خوشه‌بندی DPSC را پنج بار روی داده‌های نهایی اعمال کرده و بیشینه مقدار به دست آمده برای کیفیت خوشه‌بندی را از بین مقادیر این پنج آزمایش انتخاب کرده و به عنوان نتیجه نهایی در نظر می‌گیریم. در ادامه نتایج حاصل از الگوریتم‌هایی EDP و RMA را با هم مقایسه می‌کنیم.

مقایسه بر اساس میزان حریم خصوصی : در قضیه ۲ ثابت شد که الگوریتم پیشنهادی یا همان EDP، خاصیت حریم خصوصی تفاضلی-ϵ دارد. از طرفی در مرجع [4] اثبات شده است که الگوریتم RMA خاصیت تفاضلی- (ϵ, δ) دارد هرگاه شرط $\sigma \geq \frac{1}{\epsilon}\sqrt{10 \times \left(\epsilon + \ln\frac{1}{2\delta}\right) \times \ln\frac{n}{\delta}}$ برقرار ($m \geq 2$ و $\delta < \frac{1}{2}$) باشد. این شرط بیان می‌کند که با کاهش مقدار δ، اندازه σ به سمت بی‌نهایت میل خواهد کرد. از این‌رو، خدشه‌های با مقادیر بالا به داده‌ها اضافه خواهد شد و کارایی الگوریتم به شدت پائین خواهد آمد از این‌رو، مقدار δ هیچگاه صفر نخواهد شد. این در حالی است که مقدار δ در الگوریتم EDP، صفر است. به عبارت دیگر، در مقدار ثابت ϵ، حریم خصوصی الگوریتم ارایه شده نسبت به الگوریتم RMA بیشتر است.

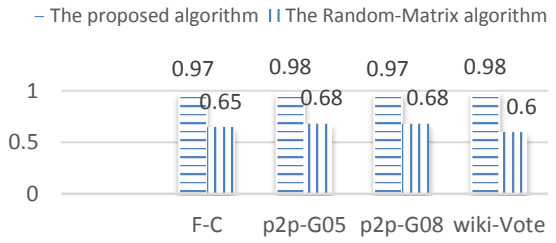
مقایسه بر اساس کیفیت خوشه‌بندی: با اعمال تبدیل موجک هار بر روی داده‌های اولیه، داده‌های نهایی دارای ابعاد به مراتب پائین‌تری خواهند بود. آزمایش‌های انجام شده در خوشه بندی‌های ۲ و ۴ هستند. معیارهای مقایسه نیز OF و NMI می‌باشند. در آزمایش‌ها، مقدار $\epsilon = 1$ و مقدار $\delta = \frac{1}{3}$ فرض شده است. شکل‌های ۵ تا ۸، نتایج مربوط به خوشه‌های ۲ و ۴ را بر اساس معیار OF بر روی داده‌های نهایی به ابعاد ۱۶ و ۱۲۸ نشان می‌دهند. همانطوری که ملاحظه می‌شود، الگوریتم پیشنهاد شده دارای مقادیر بالاتری OF نسبت به الگوریتم RMA است. همچنین، جدول‌های ۲ تا ۵، نتایج مربوط به خوشه‌های ۲ و ۴ را بر اساس معیار NMI بر روی داده‌های نهایی به ابعاد ۱۶ و ۱۲۸ نشان می‌دهند. به غیر از جدول ۵، بقیه جداول نشان می‌دهند که اندازه NMI در الگوریتم پیشنهاد شده نسبت به الگوریتم RMA بالاتر است.



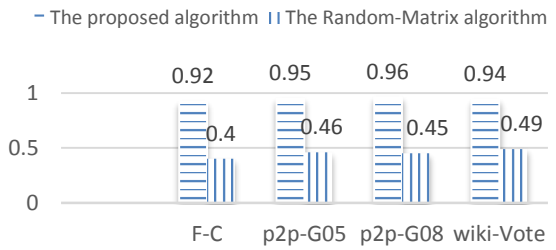
شکل ۵: ماگزیمم مقدار OF در ۵ بار آزمایش (اندازه ابعاد داده‌ی منتشر شده ۱۶ و اندازه هر خوشه ۲ است)



شکل ۶: ماگزیمم مقدار OF در ۵ بار آزمایش (اندازه ابعاد داده‌ی منتشر شده ۱۶ و اندازه هر خوشه ۴ است)



شکل ۷: ماگزیمم مقدار OF در ۵ بار آزمایش (اندازه ابعاد داده‌ی منتشر شده ۱۲۸ و اندازه هر خوشه ۲ است)



شکل ۸: ماگزیمم مقدار OF در ۵ بار آزمایش (اندازه ابعاد داده‌ی منتشر شده ۱۲۸ و اندازه هر خوشه ۴ است)

جدول (۲): ماگزیمم مقدار NMI در ۵ بار آزمایش (اندازه ابعاد داده‌ی منتشر شده ۱۶ و اندازه هر خوشه ۲ است)

Dataset	The proposed algorithm	The Random-Matrix algorithm
F-C	0.533384	0.000289
p2p-G05	0.000404	0.000283
p2p-G08	0.000464	0.000385
wiki-Vote	0.513386	0.000408

جدول۳- ماگزیمم مقدار NMI در ۵ بار آزمایش (اندازه ابعاد داده‌ی منتشر شده ۱۶ و اندازه هر خوشه۴ است)

Dataset	The proposed algorithm	The Random-Matrix algorithm
F-C	0.533384	0.000289
p2p-G05	0.000404	0.000283
p2p-G08	0.000464	0.000385
wiki-Vote	0.513386	0.000408

جدول۴- ماگزیمم مقدار NMI در ۵ بار آزمایش (اندازه ابعاد داده‌ی منتشر شده ۱۲۸ و اندازه هر خوشه۲ است)

Dataset	The proposed algorithm	The Random-Matrix algorithm
F-C	0.577343	0.000074
p2p-G05	0.000103	0.000339
p2p-G08	0.000249	0.000498
wiki-Vote	0.581113	0.000168

جدول۵- ماگزیمم مقدار NMI در ۵ بار آزمایش (اندازه ابعاد داده‌ی منتشر شده ۱۲۸ و اندازه هر خوشه۴ است)

Dataset	The proposed algorithm	The Random-Matrix algorithm
F-C	0.432104	0.000705
p2p-G05	0.000555	0.001883
p2p-G08	0.001898	0.001147
wiki-Vote	0.443203	0.000229

۷- نتیجه‌گیری

الگوریتم‌های زیادی برای حفظ حریم خصوصی یال در داده‌های منتشر شده شبکه اجتماعی ارایه شده است. در اکثر آنها، فرد مهاجم با استفاده از دانش پیش‌زمینه خود، امکان نقض حریم خصوصی داده‌ها را خواهد داشت. حریم خصوصی تفاضلی یکی از قوی‌ترین الگوریتم‌هایی است که به منظور رفع این مشکل ابداع شد. یکی از ایرادهای مهم حریم خصوصی تفاضلی، کارایی پایین آن است. در این مقاله، الگوریتم مبتنی بر حریم خصوصی تفاضلی ارایه دادیم. به‌منظور غلبه بر کارایی پایین این مفهوم، از تبدیل موجک متعامد هار، که ساده‌ترین تبدیل بین تبدیلات موجک گسسته است، استفاده نمودیم. الگوریتم پیشنهاد شده را با الگوریتمی که اخیرا ارائه شده است، بر اساس میزان حریم خصوصی و کیفیت خوشه‌بندی مقایسه نمودیم. نتایج به دست آمده نشان دادند که الگوریتم پیشنهادی هم بر اساس میزان حریم خصوصی و هم بر اساس کیفیت خوشه‌بندی، از درجه بالاتری نسبت بر الگوریتم مقایسه شده برخوردار است.

مراجع

[1] Y.Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jevon, "Analysis of topological characteristics of huge online social networking services", *In Proc. of the 16th International conference on World Wide Web, ACM*, pp. 835-844, 2007.

[2] C. Dwork, "Differential privacy: A survey of results," *in Proc. of the 5th Annual Conference on Theory and Applications of Models of Computation (TAMC'08)*, Xi'an, China, LNCS, vol. 4978. Springer-Verlag, pp. 1–19, December 2008.

[3] K.P. Soman, K.I. Ramachandran and N.G Resmi, *Insight Into Wavelets From Theory to Practice*. Prentice-Hall, 2011.

[4] F. Ahmed, R. Jin, and A.X. Liu, "A Random Matrix Approach to Differential Privacy and Structure Preserved Social Network Graph Publishing", CoRR. 2013.

[5] X. Xiao, G. Wang and J. Gehrke, "Differential privacy via wavelet transforms", *IEEE Trans. Knowl. Data Eng.* 23(8), pp. 1200-1214, 2011.

[6] Y. Wang, X. Wu, and L. Wu, "Differential privacy preserving spectral graph analysis", *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 329-340, 2013.

[7] A. Blum, C. Dwork, F. McSherry and K. Nissim, "Practical privacy: the sulq framework:, *In Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, ACM, pp. 128-138, 2005.

[8] J. Blocki, A. Blum, A. Data, and O. Sheffet, "The johnson-lindenstrauss transform itself preserves differential privacy", *In Proceedings of the 53rd IEEE Annual Symposium on Foundations of Computer Science*, IEEE, pp. 410-419, 2012.

[9] Stanford Large Network Dataset Collection ,https://snap.stanford.edu/data/, Visited: 2015-1-26.

[10] S.R.M Oliveira, O.R. Zaĭane, "Privacy-Preserving Clustering to Uphold Business Collaboration: A Dimensionality Reduction based Transformation Approach", *Int. Jounal of information Security and Privacy* (IJISP), 1(2), pp.13-36, 2007.

[11] H. Zhang, T. B. Ho, Y. Zhang and M. S. Lin, "Unsupervised feature extraction for time series clustering using orthogonal wavelet transform", *J. Informatica*, 30(3), pp. 305-319, 2006.