مجله علمی- پژوهشی "پدافند الکترونیکی و سایبری"
سال ششم، شماره ۲، تابستان ۱۳۹۷، ص ۱۰-۱

ارائه روشی برای یافتن عامل‌های پرنفوذ در انتشار اطلاعات در شبکه‌های اجتماعی مبتنی بر نظریه آنتروپی

مجید غیوری ثالث[1*]، غلامرضا بازدار[2]، ابوالفضل سرکرده‌یی[3]

۱- استادیار، ۲- دانشجوی دکتری، ۳- کارشناسی ارشد، دانشگاه جامع امام حسین(ع)

(دریافت: ۹۵/۰۴/۲۲، پذیرش: ۹۶/۰۸/۲۴)

چکیده

در حملات شناختی سایبری موضوع تحلیل انتشار اطلاعات در شبکه‌های اجتماعی برخط بسیار حائز اهمیت است که یکی از شاخه‌های اصلی در تحلیل انتشار اطلاعات یافتن عامل‌های پرنفوذ می‌باشد. در بازاریابی ویروسی نیز تحت عنوان یافتن موثرترین عامل‌ها مطرح می‌شود. در این مقاله ضمن معرفی و محاسبه دو نوع عامل پر اهمیت در انتشار اطلاعات (عامل‌های مرجع و فعال)، روشی برای یافتن این دودسته عامل‌های پراهمیت در انتشار اطلاعات در شبکه‌های اجتماعی برخط مبتنی بر نظریه آنتروپی ارائه و پیاده‌سازی شده است. روش پیشنهادی در این مقاله مبتنی است بر ارزیابی آنتروپی گراف شبکه اجتماعی برخط حاصل از انتشار اطلاعات با حذف مجموعه پرتاثیرترین عامل‌ها که بر اساس معیار درجه گره و معیار آنتروپی گره اندازه‌گیری شده است. آزمایش‌های این مقاله نشان می‌دهد که الگوریتم پیشنهادی نسبت به روش‌های قبلی، توانایی بیشتری در شناسایی مجموعه گره‌های پرنفوذ دارد، به‌طوری‌که مجموعه باقیمانده گره‌ها از همگنی قابل تنظیمی در معیار نفوذ برخوردار می‌شوند و همچنین معیاری را جهت مشخص نمودن تعداد گره‌های شاخص ارائه می‌کند.

واژه‌های کلیدی: تحلیل شبکه‌های اجتماعی برخط، انتشار اطلاعات، یافتن گره‌های پرنفوذ، نظریه آنتروپی، تحلیل گراف

۱- مقدمه

انتشار عمدی یا تصادفی اطلاعات غلط در رسانه‌های اجتماعی، به‌ویژه در موقعیت‌های حساس زمانی مانند بحران‌های جهان واقعی، می‌تواند تأثیرات مخرب جدی بر افراد و جامعه داشته باشد. به‌ویژه اگر اطلاعات مهندسی‌شده بر اساس اصول طرح‌ریزی عملیات شناختی سایبری [۱]، انتشار یابد دارای اثرات هدفمند و نفوذ در مخاطب هدف خواهد بود [۲].

امروزه می‌توان رسانه‌ها و شبکه‌های اجتماعی مجازی را به‌عنوان منابع اطلاعاتی در نظر گرفت. این شبکه‌های اجتماعی باحالت‌ها و وضعیت‌های پویایی که از خود نشان می‌دهند در قالب‌هایی مانند شکل‌گیری ارتباطات بین اعضاء شبکه، نشر خبر و بازنشر خبر، دوست‌یابی، تشکیل انجمن‌ها، انتشار محتوی، ثبت پروفایل‌های هویتی شخصی و مکانی و ... درواقع اطلاعات جدیدی را ظاهر می‌کنند. بر اساس این اقدامات و وضعیت‌ها می‌توان با دیدگاه‌های مختلف و تفسیرهای گوناگون متغیرهای تصادفی و آماره‌های مختلفی را تعریف کرد و توزیع‌های احتمال مربوطه را محاسبه کرد. بنابراین، می‌توان با به‌کارگیری مفاهیم و روابط نظریه اطلاعات، نظریه عدم قطعیت و آنتروپی پردازش‌ها و تحلیل‌های مناسبی را از این شبکه‌ها ارائه داد. به‌ویژه با توجه به پویایی و تنوع موجود در گراف شبکه‌های رسانه‌های اجتماعی مجازی، محاسبات مبتنی بر آنتروپی می‌تواند معیارخوبی برای این تنوع باشد.

در سال‌های اخیر در خصوص موضوع انتشار اطلاعات و اطلاعات غلط در شبکه‌های اجتماعی برخط، کارهای تحقیقاتی متنوعی صورت گرفته است، چراکه سرعت انتشار خیره‌کننده اطلاعات در شبکه اجتماعی و تأثیرات عمیق آن در فضای شناختی کاربران، که در مواردی موجب ایجاد تهدیدات امنیتی در سطح ملی می‌شود، پرداختن به این موضوع را ضروری نموده است [۲].

در انتشار اطلاعات یا بازاریابی در شبکه‌های اجتماعی، یک هدف عمده، حداکثرسازی تأثیر و نفوذ است. یک رویکرد اصلی برای تحقق این هدف، یافتن عامل‌های (کاربران) با بیشترین نفوذ و تأثیر در شبکه اجتماعی است. اقناع این کاربران پرنفوذ برای انتشار آن اطلاعات و یا بازاریابی محصول مورد نظر است. در جهت عکس هم برای انتشار یک اطلاعات در یک حوزه موضوعی، عوامل خاصی در شبکه اجتماعی موردنظر بیشترین نفوذ و

* رایانامه نویسنده مسئول: ghayoori@ihu.ac.ir

اثرگذاری را دارند، که با شناسایی این عامل‌ها، می‌توان توزیع اطلاعات را کنترل کرد. در این راستا کارهای تحقیقاتی متعددی به هدف یافتن پرنفوذترین عامل‌ها در شبکه اجتماعی انجام‌شده است[۶-۳]. هدف از این تحقیق، عبارت است از ارائه روشی جهت یافتن کوچک‌ترین مجموعه از عوامل پرتاثیر در انتشار اطلاعات به‌صورتی که در مابقی عامل‌های شبکه، از حیث اهمیت و نفوذ، از یکنواختی مطلوبی کمتر نباشد. روش پیشنهادی این مقاله مبتنی است بر ارزیابی اثر حذف مجموعه k گره بانفوذ بالا بر اساس معیار درجه و معیار آنتروپی گره به‌وسیله محاسبه تغییرات آنتروپی گراف شبکه اجتماعی با حذف مجموعه مذکور از گراف شبکه است. این الگوریتم، این مزیت را دارد که علاوه بر اینکه مجموعه گره‌های پرنفوذ را بر اساس اثر کلی که بر شبکه اجتماعی دارد شناسایی می‌کند، تضمین می‌کند که مجموعه گره‌های باقیمانده از همگنی قابل تنظیمی در نفوذ برخوردار هستند. ساختار ادامه مقاله به‌این‌ترتیب است: در بخش دوم مبانی نظری شامل نظریه آنتروپی در تحلیل شبکه‌های اجتماعی و تحلیل انتشار اطلاعات در شبکه‌های اجتماعی مطرح شده است، در بخش سوم کارهای تحقیقاتی مرتبط مرور شده و در بخش چهارم روش پیشنهادی ارائه شده است. در بخش پنجم نتایج آزمایش‌ها و ارزیابی نشان داده شده و بالاخره در بخش ششم جمع‌بندی و کارهای تحقیقاتی مرتبط آینده ارائه شده است.

۲- مبانی نظری و مقدمات

۲-۱- محیط نظریه آنتروپی در تحلیل شبکه‌های اجتماعی

آنتروپی را می‌توان به‌عنوان یک معیاری از محتوای اطلاعاتی مورد انتظار یا عدم قطعیت یک توزیع احتمال، تعریف کرد. همچنین آنتروپی را می‌توان به‌عنوان درجه بی‌نظمی در یک سیستم یا عدم قطعیت درباره یک بخش تعریف کرد [۷].

فرض کنید E_i نشانگر یک پیشامد باشد و p_i احتمال رخ دادن آن باشد. فرض کنید که n پیشامد $E_1,...,E_n$ با احتمال‌های $p_1,...,p_n$ است که مجموع آن‌ها یک می‌شود. ازآنجاکه رخداد وقایع با احتمال کمتر منتج به اطلاعات بیشتر می‌شود، چرا که نسبت به آن‌ها، کمترین انتظار رخداد می‌رود. بنابراین، معیار اطلاعات h یک تابع نزولی از p_i است. شانون یک تابع لگاریتمی $h(p_i)$ را برای اظهار اطلاعات پیشنهاد داد:

$$h(p_i) = log_2 \frac{1}{p_i} \qquad (۱)$$

که از بینهایت تا صفر، برای محدوده تغییر p_i از صفر تا یک کاهش می‌یابد. این تابع منعکس‌کننده این ایده است که احتمال کمتر رخداد یک پیشامد، مقدار بالاتر از اطلاعات در پیام، اظهار می‌دارد که پیشامد رخ می‌دهد .برای n مقدار اطلاعاتی $h(p_i)$، محتوای اطلاعاتی مورد انتظار، که با H نشان می‌دهیم، را آنتروپی می‌نامیم که با وزن‌دهی مقادیر اطلاعاتی توسط احتمالات مربوطه آن‌ها به دست می‌آید:

$$H = -\sum_{i=1}^{n} p_i log_2 p_i \qquad (۲)$$

تئوری آنتروپی که با توسعه تعریف بنیادی فوق شکل‌گرفته و دارای روابط و معیارهای اندازه‌گیری متنوعی مانند آنتروپی نسبی، آنتروپی شرطی، آنتروپی گراف و ... است دارای کاربردهای بسیار متنوع و وسیعی در حوزه‌های فیزیک، شیمی، ارتباطات و ... شده است. در سال‌های اخیر از آنتروپی در حل مسئله‌های مختلفی در تحلیل شبکه اجتماعی استفاده شده است مانند. اکتشاف اتصال (کشف ساختار سازمانی مخفی و انتخاب اعضاء مؤثر مدنظر مبتنی برداده شواهد نویزی و ناقص و سطح پایین است) [۵]، مسئله تشخیص انجمن [۸]، تحلیل حساسیت و مصورسازی شبکه اجتماعی [۹] .

۲-۲- تحلیل انتشار اطلاعات در شبکه‌های اجتماعی

در سال‌های اخیر در خصوص موضوع انتشار اطلاعات و اطلاعات غلط در شبکه‌های اجتماعی برخط، از جنبه‌های مختلف، کارهای تحقیقاتی متنوعی صورت گرفته است، چراکه سرعت انتشار خیره‌کننده شایعه و اطلاعات غلط در شبکه اجتماعی و تأثیرات عمیق آن در فضای شناختی کاربران، که در مواردی موجب ایجاد تهدیدات امنیتی در سطح ملی می‌شود، پرداختن به این موضوع را ضروری نموده است [۲]. در متون روانشناسی اجتماعی، یک شایعه را به‌عنوان یک روایت یا یک بیان در فضای عمومی بدون تائید یا قطعیت واقعیت‌ها تعریف می‌کنند. شایعات در فضای ابهام‌آلود، برمی‌خیزد، هنگامی‌که معنی یک موقعیت به‌آسانی مشهود نیست و یا هنگامی‌که مردم یک نیاز شدید برای امنیت را احساس می‌کنند [۱۰]. بنابراین شایعات یک نیروی قدرتمند، فراگیر و پایای مؤثر بر مردم و گروه‌هاست. انتشار شایعات و اطلاعات نادرست در مطالب مربوط به کمی‌سازی اعتبار یک قطعه اطلاعات داده شده [۱۱] و در تشخیص شیوع اطلاعات نادرست [۱۲]، مطالعه می‌شود. با رشد و رواج عمومی شبکه‌های اجتماعی برخط و پتانسیل آن‌ها در انتشار اطلاعات، قابلیت کنترل نوع اطلاعاتی که در شبکه منتشر می‌شود، خیلی مهم شده است.

تعاریف زیادی از یک شایعه وجود دارد. یک قطعه از اطلاعات در زمان گردش انتشار خود، می‌تواند بررسی‌شده باشد و می‌تواند بررسی‌شده نباشد. بعداً یک قطعه از اطلاعات که به‌عنوان درست یا غلط ارزیابی نشده باشد در حوزه‌های روانشناسی

اجتماعی به‌عنوان شایعه فرض می‌شود. البته در بعضی منابع، شایعه را فارغ از صحیح یا غلط بودن اطلاعات انتشاری، از روی معتبر و یا نامعتبر بودن منبع انتشار تعریف می‌کنند. و درواقع اطلاعات منتشرشده از یک منبع نامعتبر را شایعه می‌نامند.

اطلاعات از حیث شایعه بودن و یا نبودن را می‌توان بر اساس سه نوع ویژگی‌ها تحلیل کرد :۱- ویژگی زمانی[1] ۲- ویژگی ساختاری[2] و ۳- ویژگی زبان‌شناختی[3]

در مقابله و کنترل انتشار در شایعه، تشخیص عامل‌های پراهمیت و فعال در شیوع شایعه در شبکه‌های اجتماعی بسیار مهم است که در این مقاله روش‌هایی جهت این موضوع ارائه‌شده است.

۳- سوابق تحقیقاتی مربوطه

فعالیت‌های مسئله یافتن گره‌های پرنفوذ در انتشار اطلاعات در شبکه‌های پیچیده در [۱۶-۱۳] مطالعه شده است. برای این منظور از معیارهای مشهوری که اغلب بر اساس موقعیت گره در شبکه عمل می‌کنند استفاده‌شده است. در ادامه این معیارها مرور می‌شود.

۳-۱- روش مرکزیت درجه بالا

در این روش، گره‌های با بالاترین درجه، به‌عنوان گره‌های پرنفوذ انتخاب می‌شوند. علت این روش، این است که این گره‌ها می‌توانند تعداد بیشتری از گره‌ها را به‌طور مؤثر تحت تأثیر قرار دهد، چرا که دارای همسایگان بیشتری هستند [۱۷]. در [۱۸] این روش را توسعه داده است و معیار مرکزیت Katz معرفی می‌کند که درجه تأثیر یک گره را با به‌حساب آوردن تعداد کل گام‌ها اندازه‌گیری می‌کند. هر اتصال با فاصله j با β^j جریمه می‌شود که $1 \geq \beta \geq 0$. رابطه محاسبه این معیار به‌صورت زیر است:

$$C_i^{Katz} = e_i^T \left(\sum_{j=1}^{\infty} (\beta \mathbf{A})^j \right) \mathbf{I} \qquad (۳)$$

که e_i یک بردار ستونی است که همه درایه‌هایش صفر است جز درایه i ام که ۱ است. و I ماتریس همانی است. معیار مرکزیت درجه بالا دارای این نقص است که یک گره را به‌صورت محلی در نظر می‌گیرد و به‌عبارت‌دیگر مبتنی بر موقعیت محلی گره است و در گراف‌هایی با مؤلفه‌های متعدد، احتمالاً گره‌های مؤثر فقط از یک مؤلفه بزرگ گراف انتخاب می‌شود و اشکال عمده دیگر این است که با این معیار اثر گره در لایه‌های بعدی ارتباطی و در کل شبکه دیده نشده است .

۳-۲- روش مرکزیت نزدیکی

معیار دوری[4] یک گره، به‌صورت مجموع فاصله گره از همه گره‌های دیگر محاسبه می‌شود و معیار نزدیکی را با معکوس معیار دوری محاسبه می‌شود. بنابراین، معیار نزدیکی را می‌توان به‌عنوان معیاری که نشان می‌دهد که چقدر طول می‌کشد اطلاعات از یک گره به همه گره‌ها به‌صورت متوالی منتشر شود، در نظر گرفت. به‌عبارت‌دیگر، بر اساس راهبرد نزدیکی، یک گره پرنفوذ (مرکزی) است اگر فاصله‌اش از همه گره‌های دیگر کمترین باشد. این گره‌ها به‌واسطه اینکه کمترین تعداد واسطه رادارند، نفوذ بیشتری دارند. این معیار مرکزیت می‌توان با شمردن کوتاه‌ترین مسیرها محاسبه شود. و در زیر عبارت معروف اشاره می‌شود [۱۹]:

$$C_i^{CLO} = e_i^T SI \qquad (۴)$$

که S ماتریسی است که درایه (i, j) آن، طول کوتاه‌ترین مسیر بین i و j را نشان می‌دهد. معیار نزدیکی به گردش روی کل شبکه احتیاج دارد و کاملا مشهود است برای شبکه با مقیاس بزرگ زمان‌بر و نامناسب است.

۳-۳- روش مرکزیت مابینی

با این شاخص، گره‌های مؤثر آن‌هایی هستند که در بیشترین تعداد کوتاه‌ترین مسیرها از همه گره‌ها به گره‌های دیگر رؤیت می‌شوند. در [۱۹] رابطه زیر برای این تعریف ارائه‌شده است:

$$C_i^{BET} = \sum_{j \neq r \neq i} \frac{g_{ir(i)}}{g_{ir}} \qquad (۵)$$

که g_{ir}، تعداد کوتاه‌ترین مسیرها بین گره‌های j و r است و $g_{ir(i)}$، تعداد کوتاه‌ترین مسیرهای بین j و r است که از گره i عبور می‌کند. گره‌های با بالاترین مابینی را تنگه[5] [۲۰] یا میانجی[6] [۲۱] یا حفره‌های ساختاری[7] [۲۲] گویند.

۳-۴- روش مرکزیت بردار ویژه

در این روش که در [۲۳] معرفی‌شده است و ارتباط نزدیکی با معیار مرکزیت Katz دارد، سعی می‌کند که تأثیر و نفوذ یک گره را با تخصیص یک امتیاز به هر گره مبتنی بر مجاورت آن گره با گره‌های با امتیازبندی بالا، پیدا کند.

1- Temporal Characteristic
2- Structural Characteristic
3- Linguistic Characteristic

4 -Farness
5 -Bottlenecks
6 -Intermediaries
7- Structural Holes

۳-۵- روش رتبه صفحه

رتبه صفحه، الگوریتمی است که در موتور جستجوی گوگل برای رتبه‌بندی صفحات وب استفاده می‌شود. یک صفحه وب که به صفحات وب مهم‌تری لینک شده باشد، رتبه بالاتری را دارد. بنابراین یک صفحه با همسایگان کمتر ممکن است یک رتبه صفحه بالاتری نسبت به صفحه با همسایگان بیشتر داشته باشد. در [۲۵-۲۴] از این معیار مرکزیت برای رتبه‌بندی گره‌ها در شبکه اجتماعی استفاده‌شده است.

۳-۶- مرکزیت تنزیل درجه

در[۲۶]، الگوریتم اکتشافی تنزیل درجه ارائه شده است. هنگامی‌که یک گره به‌عنوان یک گره اولیه انتخاب‌شده است، گره دیگری با بیشترین درجه را می‌توان بالقوه به‌عنوان گره اولیه جدید در نظر گرفت، اما یال بین این دو گره را نباید در درجه‌اش حساب کرد [۲۷]. به‌عبارت دیگر اگر یک گره u درجه d_u، و d'_u از آن‌ها قبلاً به‌عنوان گره اولیه انتخاب‌شده باشد، احتیاج است که $d(u)$ با $2d'_u+(d_u-d'_u)d'_u p$ تخفیف داده می‌شود، که p یک احتمال انتشار کوچک است. این مدل جریان اطلاعات کل در شبکه را بیشینه نمی‌کند.

۳-۷- روش رتبه سردسته[1]

در [۲۸]، یک نوع متفاوت از رتبه صفحه تحت عنوان رتبه سردسته ارائه شده است. روش رتبه سردسته وزن‌دار که نسخه به‌خوبی بهبودیافته از روش رتبه سردسته است در [۲۹] ارائه شده است.

۳-۸- تجزیه k قشر[2]

در [۳۰]، این معیار ارائه‌شده است که بر اساس موقعیت گره در شبکه عمل می‌کند و شاخص k_s را به هر گره تخصیص می‌دهد. گره‌های با شاخص بالا در اعماق هسته شبکه واقع شده‌اند و گره‌های با شاخص پایین در پیرامون شبکه واقع شده‌اند.

۳-۹- روش الگوریتم حریصانه

در این الگوریتم که در [۳۱] ارائه شده است، یک مجموعه از گره‌های اولیه، S، مفروض است و در هر مرحله از الگوریتم یک گره تنها، v، به مجموعه S افزوده می‌شود به‌طوری‌که $S \cup \{v\}$ گسترش تأثیر را بیشینه کند و تعداد زیادتری از گره‌ها را در شبکه فعال می‌کند. این فرایند به‌طور تکراری ادامه می‌یابد تا k گره برتر انتخاب شود. به‌عبارت‌دیگر $|S|=k$.

همان‌طور که در بالا ارائه شد، این روش‌ها، اثر تجمعی هر گره را در کل شبکه و همچنین مجموع اثرات تجمعی مهم‌ترین گره‌ها را در کل شبکه لحاظ نکرده‌اند. در روش پیشنهادی ما، ضمن اینکه از خاصیت تجمعی رابطه آنتروپی برای سنجش میزان اثرگذاری مجموعه گره‌های منتخب استفاده شده است، با به‌کارگیری پارامترهای آستانه‌ای، مجموعه منتخب گره‌ها را تا حد مورد نظر نسبت به بقیه شبکه متمایز انتخاب می‌کند.

۴- روش پیشنهادی

شناسایی و مدل سازی رفتار عامل های موجود در شبکه اجتماعی موردنظر در مطالعه فرایند انتشار اطلاعات و به‌ویژه انتشار و شیوع شایعه، ضروری است. این رفتار در شبکه اجتماعی توییتر (که موردمطالعه این مقاله است) به‌صورت رفتارهایی مانند نشر خبر، بازنشر خبر، پاسخ دادن به خبر، ذکر کردن فرد در خبر و دنبال شدن فرد صورت می‌گیرد. در این مقاله به این نکته توجه شده که نشر خبر کردن یک اقدام فعالانه است و اگر پیام عامل الف توسط عامل ب بازنشر شود این نشانه‌ای از مرجع بودن عامل الف و سطحی از فعال بودن توسط عامل ب است اگرچه اقدام نشر پیام از اقدام بازنشر پیام، اقدامی فعالانه‌تری ارزیابی می‌شود. اقدام پاسخ دادن به خبر به یک پیام سطح بالاتری از واکنش است. اگر هم فرد الف از ب در پیام خود اشاره کند، نشانی از اهمیت مرجع بودن فرد ب است. به‌این‌ترتیب ما مجموعه اقدامات بازنشر خبر، پاسخ دادن به خبر و ذکر کردن فرد در خبر را برای یک عامل نشان میزان فعال بودن عامل می‌دانیم و مجموعه بازنشر خبر یک عامل، پاسخ دادن به خبر یک عامل و ذکر شدن یک عامل در خبر را میزان مرجع بودن آن عامل در نظر می‌گیریم. حال اگر خواسته شود میزان اهمیت عامل‌ها را از حیث فعال بودن و مرجع بودن را در فرایند انتشار یک شایعه محاسبه شود، باید سنجه‌های مناسبی برای این موضوع انتخاب شود. در ادامه روش‌های مختلف پیشنهادی محاسبه میزان اهمیت عامل‌های موجود در شبکه اجتماعی توییتر در انتشار شایعه ارائه می‌شود:

الف : درروش اول ، درجه ورودی $D_{in}(i)$ و درجه خروجی $D_{out}(i)$ همه گره‌ها را به ترتیب زیر محاسبه می‌کنیم:

$$D_{in}(i)=D_{in}^{rt}(i)+D_{in}^{re}(i)+D_{in}^{mt}(i) \qquad (۶)$$

$$D_{out}(i)=D_{out}^{rt}(i)+D_{out}^{re}(i)+D_{out}^{mt}(i) \qquad (۷)$$

به‌سادگی دیده می‌شود که $D_{in}(i)$ معیاری از خصوصیت مرجعیت در انتشار شایعه است و بنابراین می‌توان آن را با $D_{ref}(i)$ نشان داد و $D_{out}(i)$ معیاری از فعال بودن عامل در انتشار شایعه است و بنابراین می‌توان آن را با $D_{act}(i)$ است. از روی این دو آماره می‌توانیم توزیع احتمال معیار مشهوریت

1- Leader Rank
2- K-Shell Decomposition

$P_{in}(i)$ و توزیع احتمال معیار جمع‌گرایی $P_{out}(i)$ عامل‌های شبکه را محاسبه کرد:

$$P_{ref}(i)=\frac{D_{ref}(i)}{\sum_{j=1}^{N}D_{ref}(j)} \quad (۸)$$

$$P_{act}(i)=\frac{D_{act}(i)}{\sum_{j=1}^{N}D_{act}(j)} \quad (۹)$$

با تعریف توزیع‌های احتمال فوق، می‌توان عامل‌های موجود در شبکه را از حیث مرجعیت در انتشار اطلاعات یا فعال بودن در انتشار اطلاعات مقایسه شود، و درجه اهمیت در انتشار اطلاعات را از این دو حیث مقایسه و پراهمیت‌ترین عامل‌ها را از حیث مرجعیت انتشار و فعالیت در انتشار شناسایی شود. از طرفی می‌توان کل نقش هر عامل را در انتشار اطلاعات با محاسبه تابع توزیع کل برای هر عامل محاسبه شود:

$$D(i)=D_{ref}(i)+D_{act}(i) \quad (۱۰)$$

$$P(i)=P_{ref}(i)+P_{act}(i) \quad (۱۱)$$

ب: در شناسایی عامل‌های شاخص و پراهمیت در انتشار اطلاعات در شبکه اجتماعی بر اساس عامل‌های با درجه بالاتر، اثر هر عامل در وضعیت فعالیتی دیگر عامل‌ها و کل شبکه دیده نمی‌شود، لذا با استفاده از مفاهیم نظریه آنتروپی می‌توان این نقیصه را رفع نمود. در این روش آنتروپی شبکه برای سه وضعیت: ۱.شبکه کل فعالیت‌های مرجعیت و فعال بودن عامل‌ها در انتشار، ۲. شبکه فعالیت مرجعیت عامل‌ها در انتشار، ۳. شبکه فعالیت فعال بودن عامل‌ها در انتشار را محاسبه و نرمال می‌شود. سپس با حذف هر گره در هر سه شبکه و محاسبه آنتروپی نرمال شده، می‌توان عامل‌هایی که با حذفشان بیشترین یکنواختی در وضعیت فعالیتی دیگر عامل‌ها ایجاد می‌شود را به‌عنوان عامل‌های پراهمیت معرفی کرد.

$$H_{ref}=-\sum_{i=1}^{N}\mathrm{p}_{ref}(\mathrm{i})\log_2\mathrm{p}_{ref}(\mathrm{i}) \quad (۱۲)$$

$$H_{act}=-\sum_{i=1}^{N}\mathrm{p}_{act}(\mathrm{i})\log_2\mathrm{p}_{act}(\mathrm{i}) \quad (۱۳)$$

$$Ru_{ref}=\frac{H_{ref}}{\log_2 N} \quad (۱۴)$$

$$Ru_{act}=\frac{H_{act}}{\log_2 N} \quad (۱۵)$$

$$H_{ref}(i)=-\sum_{j=1,j\neq i}^{N-1}\mathrm{p}_{ref}(j)\log_2\mathrm{p}_{ref}(j) \quad (۱۶)$$

$$H_{act}(i)=-\sum_{j=1,j\neq i}^{N-1}\mathrm{p}_{act}(j)\log_2\mathrm{p}_{act}(j) \quad (۱۷)$$

$$Ru_{ref}(i)=\frac{H_{ref}(i)}{\log_2(N-1)} \quad (۱۸)$$

$$Ru_{act}(i)=\frac{H_{act}(i)}{\log_2(N-1)} \quad (۱۹)$$

ج: درروش دوم ارزش درجه بالا بودن عامل‌ها در تعیین عامل‌های پراهمیت در انتشار شایعه دیده نشده است. برای بهره‌گیری از فواید هر دو روش فوق و همچنین یافتن حدی برای تعداد پراهمیت‌ترین عامل‌های انتشار شایعه، از روش ارائه‌شده در [۳۲] که به هدف نمایه‌سازی ترافیک شبکه کامپیوتری استفاده‌شده است، الهام گرفته شده و روشی مبتنی بر الگوریتم زیر پیاده‌سازی و آزمایش شد.

همان‌طور که در الگوریتم ارائه‌شده دیده می‌شود، در این روش مجموعه S از عامل‌های با بیشترین احتمال و درجه را به‌طوری پیدا می‌کنیم که با حذف آن از عامل‌های شبکه، مقدار یکنواختی و همسانی احتمالات عامل‌های باقیمانده، به‌قدر کافی و دلخواه ما شود. برای تحقق مطلوب ما، مقدار یکنواختی و همسانی احتمالات عامل‌های باقیمانده را از طریق محاسبه آنتروپی نرمال‌شده (Ru) عامل‌های باقیمانده محاسبه می‌شود و دو پارامتر α و β تعریف می‌شود که α حد آستانه مقدار احتمال عامل‌های مجموعه S را تعیین می‌کند و β حد مطلوب RU عامل‌های باقیمانده را نشان می‌دهد.

1: Parameters: $\alpha := \alpha_0; \beta := 0.9; S := \emptyset;$
2: Initialization: $S := \emptyset; R := A; K := 0;$
3: compute probe.dist.P_R and its RU $\theta := RU(\mathrm{P}_R)$
4: while $\theta \leq \beta$ do
5: $\alpha = \alpha \times 2^{-k}; k++;$
6: for each $a_i \in R$ do
7: if $\mathrm{P}_A(a_i) \geq \alpha$ then
8: $S := S \cup \{a_i\}; R := R - \{a_i\};$
9: end if
10: end for
11: compute(cond.) prob.dist. P_R and $\theta := RU(\mathrm{P}_R);$
12: end while

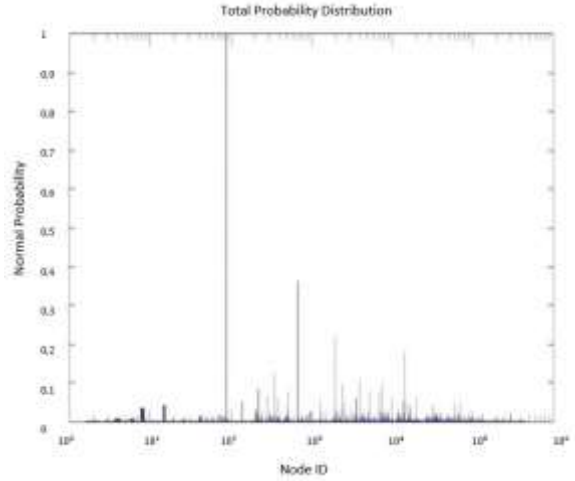
این الگوریتم برای سه وضعیت شبکه فعالیت مرجعیت انتشار شایعه عامل‌ها و شبکه فعالیت فعال بودن در انتشار شایعه عامل‌ها و وضعیت شبکه کل فعالیت‌های عامل‌ها در شبکه پیاده‌سازی و آزمایش‌ها لازم صورت گرفته است. در ادامه نتایج پیاده‌سازی هر سه روش و آزمایش‌ها مربوطه تشریح شده است.

۵- نتایج آزمایش‌ها و ارزیابی

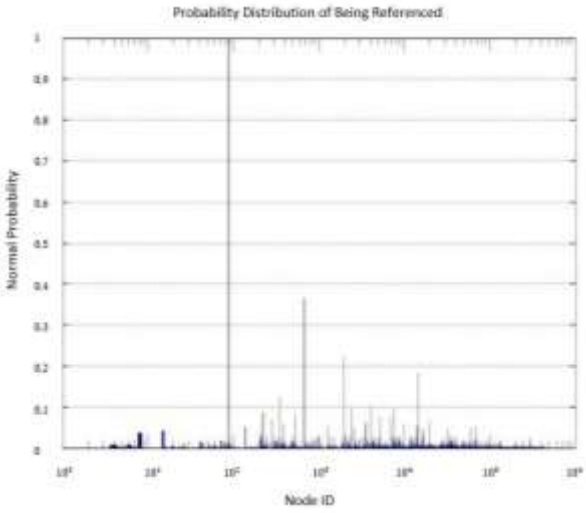
برای ارزیابی روش پیشنهادی از مجموعه داده Higgs Twitter که در خصوص انتشار شایعه علمی در شبکه اجتماعی توییتر جمع‌آوری‌شده استفاده شد [۳۳]. این مجموعه داده توسط دانشگاه استنفورد تهیه شده است. در این مجموعه داده ۴ گراف جهت‌دار از پیام‌های ارسال‌شده از تاریخ اول تا هفتم جولای ۲۰۱۲ به شرح زیر تهیه شده است. در تاریخ ۴ جولای ۲۰۱۲ خبری در خصوص Higgs boson که در حوزه ذره بنیادی اولیه است منتشر شد که بر این اساس نشر خبرهایی که در آن مجموعه کلمات کلیدی مرتبط با این موضوع در شبکه اجتماعی توییتر منتشر شده‌اند، قبل از این تاریخ و بعد از آن و همچنین در حین آن در قالب فعالیت کاربران و بر اساس کلمات کلیدی مشخص ثبت شده است. در مجموعه گراف‌ها ID کاربران به‌صورت گمنام و در قالب انتساب عدد به کاربران آمده است که در چهار گراف این اعداد برای یک ID یکسان است و می‌توان به‌صورت ترکیبی از این گراف‌های جهت‌دار استفاده کرد. در ادامه چهار گراف جهت‌دار این مجموعه داده معرفی می‌شود. در این مجموعه داده کلیه فعالیت‌های عامل در قالب مجموعه داده فعالیت که شامل سه دسته فعالیت برای هر عامل هست آمده است که در این مقاله مبتنی بر آن کلیه روش‌ها ارزیابی‌شده است. پیاده‌سازی الگوریتم پیشنهادی و اجرای آزمایش‌های در بستر نرم‌افزار MATLAB صورت گرفته است.

۵-۱- پیاده‌سازی روش اول

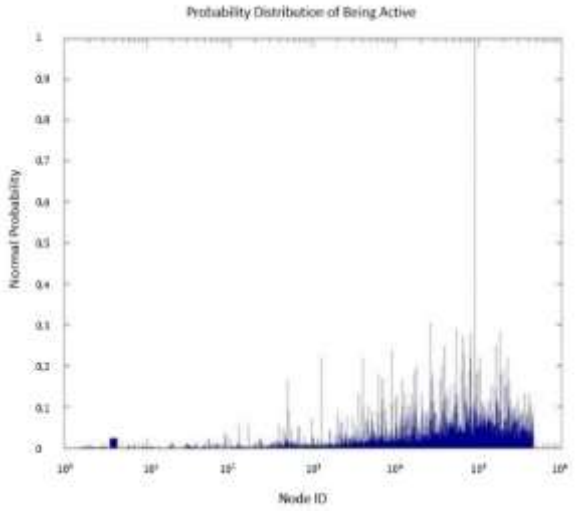
مطابق رابطه (۱۱) توزیع احتمال کل شبکه محاسبه و مطابق شکل (۱) نمایش داده‌شده است. به‌این‌ترتیب عامل‌های با بیشترین درجه و احتمال فعالیت در انتشار شایعه قابل تشخیص خواهند بود. مطابق رابطه (۹) توزیع احتمال شبکه فعالیت مرجعیت محاسبه و مطابق شکل (۲) نمایش داده شده است. به‌این‌ترتیب عامل‌های با بیشترین درجه و احتمال فعالیت مرجعیت در انتشار شایعه قابل‌تشخیص خواهند بود. مطابق رابطه (۸) توزیع احتمال شبکه فعالیت فعال بودن محاسبه و مطابق شکل (۳) نمایش داده‌شده است. به‌این‌ترتیب عامل‌های با بیشترین درجه و احتمال فعالیت فعال‌بودن در انتشار شایعه قابل‌تشخیص خواهند بود.



شکل (۱): توزیع احتمال کل



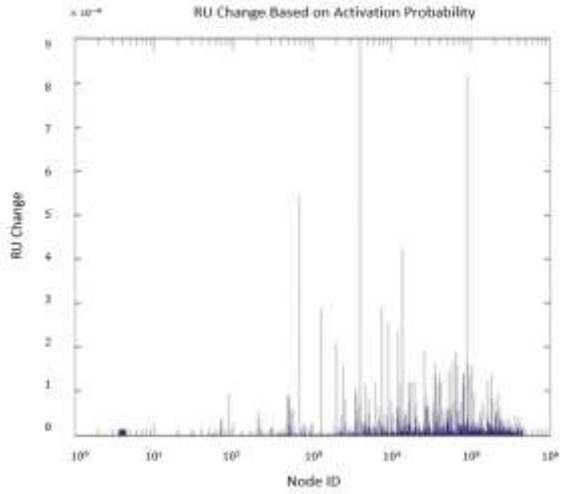
شکل (۲): توزیع احتمال مرجعیت



شکل (۳): توزیع احتمال فعالیت

۵-۲- پیاده‌سازی روش دوم

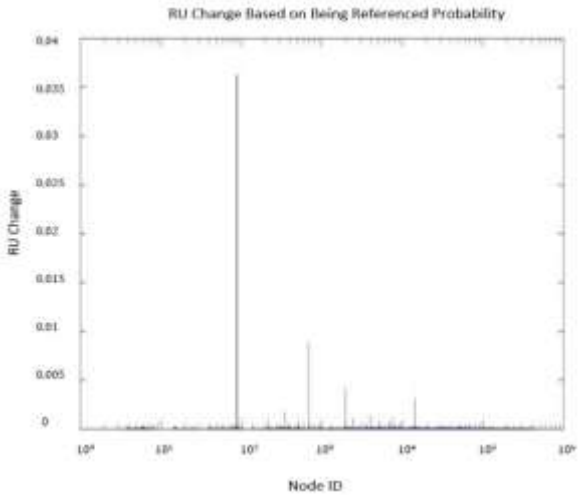
مطابق رابطه (۱۵) تغییرات آنتروپی مبتنی بر احتمال فعالیت با حذف هر گره محاسبه و مطابق شکل (۴) نمایش داده‌شده است. به‌این‌ترتیب عامل‌های فعال با بیشترین اثر سراسری در انتشار شایعه قابل تشخیص خواهند بود.



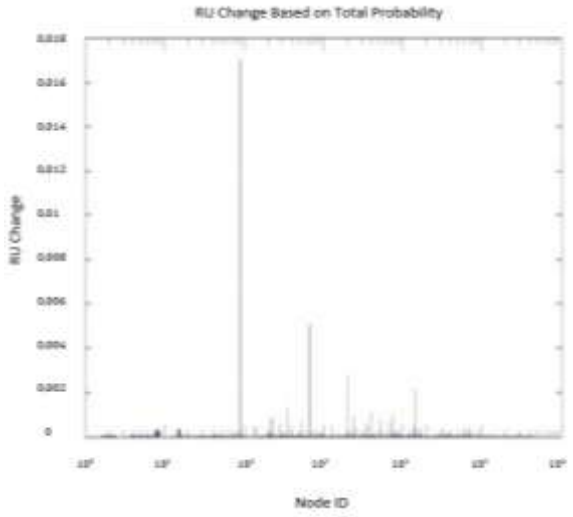
شکل (۴): تغییرات آنتروپی مبتنی بر احتمال فعالیت با حذف هر گره

مطابق رابطه (۱۴) تغییرات آنتروپی مبتنی بر احتمال مرجعیت با حذف هر گره محاسبه و مطابق شکل (۵) نمایش داده‌شده است. به‌این‌ترتیب عامل‌های مرجع با بیشترین اثر سراسری در انتشار شایعه قابل‌تشخیص خواهند بود.

همچنین تغییرات آنتروپی مبتنی بر احتمال کل با حذف هر گره محاسبه‌شده است که در شکل (۶) ارائه شده است.



شکل (۵): تغییرات آنتروپی مبتنی بر احتمال مرجعیت با حذف هر گره



شکل (۶): تغییرات آنتروپی مبتنی بر حذف گره در کل گره‌ها

۵-۳- پیاده‌سازی روش پیشنهادی

درروش سوم مبتنی بر الگوریتم ارائه‌شده برای سه دسته عامل مرجع، فعال و کل، عامل‌های شاخص احصاء شده است که نتایج حاصله به شرح زیر هست.

الف) برای عامل‌های مرجع:

در این دسته تعداد ۵۶۴۸ عامل وجود دارد که در الگوریتم ارائه‌شده میزان α برابر با ۰/۰۰۶۳ بیشترین مقدار احتمال و β برابر با ۰/۹ در نظر گرفته‌شده است.

جدول (۱): گره‌های شاخص در بین گره‌های مرجع

ردیف	روش مورد استفاده	تعداد نودهای شاخص بدست آمده	نام کاربری نودهای شاخص به ترتیب
۱	مبتنی بر حد آستانه احتمالی	۵	ID=88,677,1988,14454,349
۲	مبتنی بر حد آستانه تغییرات آنتروپی نرمال شده	۸	ID=88,677,1988,14454,349,3998,7533,2417
۳	مبتنی بر روش ارائه شده	۱۸۹	ID=88,677,220,349,1988,2417,3998,7533,14454,8,15,42,43,71,77,81,138,205,206,213,216,236,245,248,268,280,283,301,310,317 ۳۰ نود شاخص ابتدایی

ب) عامل‌های فعال:

در این دسته تعداد ۴۸۵۹۰ عامل وجود دارد که در الگوریتم ارائه‌شده میزان α برابر با ۰/۰۰۱۱ بیشترین مقدار احتمال و β برابر با ۰/۹۸در نظر گرفته‌شده است.

جدول (۲): گره‌های شاخص در بین گره‌های فعال

ردیف	روش مورد استفاده	تعداد نودهای شاخص بدست آمده	نام کاربری نودهای شاخص به ترتیب
۱	مبتنی بر حد آستانه احتمالی	۱	ID=89805
۲	مبتنی بر حد آستانه تغییرات آنتروپی نرمال شده	۱	ID=89805
۳	مبتنی بر روش ارائه شده	۳۱۹	ID=89805,9021,26158,38535,53508,64911,81181,81405,162402,181190 ۱۰ نود شاخص ابتدایی

ج) کل عامل‌ها:

در این دسته تعداد ۴۷۱۰۱ عامل وجود دارد که در الگوریتم ارائه‌شده میزان α برابر با ۰/۰۰۳۱ بیشترین مقدار احتمال و β برابر با ۰/۹در نظر گرفته‌شده است.

جدول (۳): گره‌های شاخص در بین کل گره‌ها

ردیف	روش مورد استفاده	تعداد نودهای شاخص بدست آمده	نام کاربری نودهای شاخص به ترتیب
۱	مبتنی بر حد آستانه احتمالی	۵	ID=88,677,1988,14454,349
۲	مبتنی بر حد آستانه تغییرات آنتروپی نرمال شده	۶	ID=88,677,1988,14454,349, 3998
۳	مبتنی بر روش ارائه شده	۱۳۴	ID=88,677,1988,14454,8,15,71,77,138,205,206,213,216,220,236,245,280,283,301,317,322,335,356,383,485,503,511,519,587 ۳۰ نود شاخص ابتدایی

۵-۴- ارزیابی

در روش‌های قبلی روشی برای تعیین تعداد گره‌های شاخص ارائه نشده است که این امر موجب می‌شود نتوان تعیین کرد چه تعداد گره شاخص در یک شبکه باید انتخاب شود. همچنین در روش‌های قبلی معمولاً مبتنی بر محاسبات محلی گره‌های شاخص تعیین شده‌اند.

در این مقاله باهدف اینکه بتوانیم دو موضوع محلی و سراسری بودن، انتخاب مجموعه گره‌های شاخص و تعیین مرز تعداد گره‌های شاخص را بررسی نماییم دو روش از روش‌های قبلی و یک روش پیشنهادی پیاده‌سازی شده است. روش اول که بر اساس درجه گره است به‌صورت محلی بوده و مرز تعداد گره‌های شاخص را تعیین نمی‌کند. روش دوم که مبتنی بر تغییرات آنتروپی با حذف تک گره است به‌صورت سراسری بوده ولی همچنان مرز تعداد گره‌های شاخص را تعیین نمی‌کند. براین اساس روشی پیشنهاد شده است که مبتنی بر آنتروپی می‌باشد که یک معیاری برای ارزیابی گره در کل شبکه است که به همین دلیل ارزیابی گره به‌صورت سراسری است و با وجود معیار β امکان تعیین گره‌های شاخص مبتنی بر نیاز مسئله فراهم می‌شود.

جهت ارزیابی روش پیشنهادی در سه جدول این روش با دو روش دیگر مقایسه شده است در روش اول و دوم معیاری برای تعیین مرز تعداد گره‌های شاخص ارائه نشده است به این دلیل فقط می‌توان گره‌هایی را شاخص در نظر گرفت که اندازه معیار آن‌ها بسیار متمایز از دیگر گره‌ها است و به همین دلیل گره‌های شاخص به‌صورت محدود مشخص شده‌اند و همان‌طور که در جدول‌های (۱ و ۲) مشاهده می‌شود در این دو روش نسبت به روش پیشنهادی تعداد بسیار کمتری گره شاخص تشخیص داده شده است و اگر میزان فعالیت گره‌ها به هم نزدیک باشد نمی‌تواند گره‌های شاخص را به‌درستی تعیین نماید مانند جدول (۲) که در آن مبتنی بر این دو روش یک گره شاخص تشخیص داده شده است. درصورتی‌که روش پیشنهادی با هدف ارائه یک مرز جهت تعیین گره‌های شاخص ارائه شد. بر این اساس در جدول (۲)، ۳۱۹ گره شاخص در بین گره‌های فعال مشخص‌شده است.

همان‌طور که در نتایج روش پیشنهادی قابل مشاهده است تعداد گره‌های شاخص فعال به‌طور قابل‌توجهی از گره‌های شاخص مرجع بیشتر است.

۶- نتیجه‌گیری و کارهای آینده

در روش پیشنهادی معیاری جهت تعیین تعداد گره‌های شاخص ارائه شد. این روش امکان ارزیابی کامل گره‌های شاخص فعال و مرجع در انتشار یک شایعه را فراهم می‌کند. همچنین امکان مقایسه وضعیت نسبی انتشاری شایعه های مختلف را فراهم می‌آورد.

در ادامه کار می‌توان بروی محورهای زیر تمرکز کرد:

- ارائه ویژگی‌های شناختی عامل‌های انتشار اطلاعات بر اساس نظریه آنتروپی

- ارائه روشی جهت دخیل کردن دنبال کننده‌های هر عامل در تشخیص عامل‌های شاخص در انتشار اطلاعات
- روشی جهت اولویت‌بندی عامل شاخص با توجه به تعداد زیاد عامل‌ها در شبکه‌های اجتماعی مجازی پس از تعیین عامل‌های شاخص

۷- مراجع

[1] Gh. Bazdar, and H. Lashgarian, "Cyber Cognitive Warfare Recognition Based on Cognitive Science", 2nd National Conference on Cyber ۏefence, Emam Hossein University, Tehran, 2013. (In Persian).

[2] S. Vosoughi, "Automatic Detection and Verification of Rumors on Twitter," no. 2008, pp. 1–147, 2015.

[3] D. Kempe, J. Kleinberg, and É. Tardos, "Influential Nodes in a Diffusion Model for Social Networks," Autom. Lang. Program., vol. 3580, pp. 1127–1138, 2005.

[4] C. Aggarwal, S. Lin, and P. S. Yu, "On Influential Node Discovery in Dynamic Social Networks."

[5] J. Shetty and J. Adibi, "Discovering Important Nodes through Graph Entropy: The Case of Enron Email Database," Proc. 3rd Int. Work. Link Discov., pp. 74–81, 2005.

[6] "Community-based Greedy Algorithm for Mining Top-K Influential Nodes in Mobile Social Networks Categories and Subject Descriptors," pp. 1039–1048.

[7] C. E. Shannon and W. Weaver, "The Mathematical Theory of Communication," Math. theory Commun., vol. 27, no. 4, p. 117, 1949.

[8] J. D. Cruz, C. Bothorel, and F. Poulet, Entropy based community detection in augmented social networks. 2011, pp. 163–168.

[9] E. Serin and S. Balcisoy, "Entropy Based Sensitivity Analysis and Visualization of Social Networks," 2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., pp. 1099–1104, 2012.

[10] N. DiFonzo and P. Bordia, Rumor Psychology: Social and Organizational Approaches, vol. x. 2007.

[11] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," Proc. 20th Int. Conf. World wide web - WWW '11, p. 675, 2011.

[12] J. Ratkiewicz, M. D. Conover, M. Meiss, B. Gonc, A. Flammini, and F. Menczer, "Detecting and Tracking Political Abuse in Social Media," Artif. Intell., pp. 297–304, 2011.

[13] A. Shokrollahi, "Improving detection of influential nodes in complex networks," 2015.

[14] Z. Zhao, X. Wang, W. Zhang, and Z. Zhu, "A Community-Based Approach to Identifying Influential Spreaders," pp. 2228–2252, 2015.

[15] R. Narayanam, "A Shapley Value Based Approach to Discover Influential Nodes in Social Networks," pp. 1–18.

[16] M. G. Rossi, F. D. Malliaros, and M. Vazirgiannis, "Spread It Good , Spread It Fast : Identification of Influential Nodes in Social Networks," pp. 101–102.

[17] J. Golbeck, Analyzing the Social Web. 2013.

[18] L. Katz, "A new status index derived from sociometric analysis," Psychometrika, vol. 18, no. 1, pp. 39–43, 1953.

[19] L. C. Freeman, "Centrality in social networks conceptual clarification," Soc. Networks, vol. 1, no. 3, pp. 215–239, 1978.

[20] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, "The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics," PLoS Comput. Biol., vol. 3, no. 4, pp. 713–720, 2007.

[21] P. J. Carrington and J. Scott, The SAGE Handbook of social network analysis. Introduction. 2011.

[22] R. S. Burt, Structural holes: The social structure of competition. 1995.

[23] P. Bonacich and P. Lloyd, "Eigenvector-like measures of centrality for asymmetric relations," Soc. Networks, vol. 23, no. 3, pp. 191–201, 2001.

[24] S. Fortunato, M. Boguna, A. Flammini, and F. Menczer, "How to make the top ten: Approximating PageRank from in-degree," Arxiv Prepr. cs/0511016, p. 8, 2005.

[25] J. Heidemann, M. Klier, and F. Probst, "Identifying key users in Online Social Networks: A PageRank Based Approach," Proc. 31st Int. Conf. Inf. Syst., vol. 4801, no. December, pp. 1–22, 2010.

[26] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," Time, vol. 67, no. 1, p. 199, 2009.

[27] C. C. Aggarwal, "An Introduction to social network data analytics," in Social Network Data Analytics, 2011, pp. 1–15.

[28] L. Lü, Y.-C. Zhang, C. H. Yeung, and T. Zhou, "Leaders in Social Networks, the <italic>Delicious</italic> Case," PLoS One, vol. 6, no. 6, p. e21202, 2011.

[29] Q. Li, T. Zhou, L. Lü, and D. Chen, "Identifying influential spreaders by weighted LeaderRank," Phys. A Stat. Mech. its Appl., vol. 404, pp. 47–55, 2014.

[30] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. a. Makse, "Identifying influential spreaders in complex networks," Nat. Phys., vol. 6, no. 11, p. 36, 2010.

[31] D. Kempe and J. Kleinberg, "Maximizing the Spread of Influence through a Social Network," 2003.

[32] K. Xu, Z. L. Zhang, and S. Bhattacharyya, "Internet Traffic Behavior Profiling for Network Security Monitoring," IEEE/ACM Transactions on Networking, 2008.

[33] M. De Domenico, A. Lima, P. Mougel, and M. Musolesi, "The anatomy of a scientific rumor.," Sci. Rep., vol. 3, p. 2980, 2013.

Journal of Electronical & Cyber Defence
Vol. 6, No. 2, 2018, Serial No. 22

Introduction of the Entropy-Based Method for Finding Influential Nodes in Information Dissemination on Online Social Networks

M. Ghayoori Sales*, Gh. R. Bazdar, A. Sarkardei

*Imam Hossein University

(Received: 12/07/2016, Accepted: 15/11/2017)

ABSTRACT

In cognitive cyber-attacks, information dissemination analysis in online social networks is a very important issue. One of the main branches of information dissemination analysis is finding influential nodes which this issue is also arisen in viral marketing as finding the influential individuals. In this paper, while introducing and calculating two types of important nodes in information dissemination (reference and active nodes), a method for identifying these two important types of nodes in the dissemination of information in online social networks is presented and implemented based on the entropy theory.
The proposed method in this paper is based on the evaluation of the entropy of the online social network graph generated from dissemination of information by removing the set of the most influential nodes measured on the basis of the nodal-degree and the entropy of the nodes. The experiments of this paper show that the proposed algorithm is more capable of identifying the set of influential nodes than the previous methods, in a way that the remaining set of nodes will have an adjustable homogeneity in influence measure and also presents a measure for determining the number of influencer nodes.

Keywords: Social Network Analysis, Information Dissemination, Influential Nodes Finding, Entropy Theory, Graph Analysis

*Corresponding Author Email: ghayoori@ihu.ac.ir