

# یادگیری احتمالات انتشار برای مدل آبخاری مستقل

یوسف مشایخی<sup>۱</sup>، محمدرضا میبیدی<sup>۲</sup>

<sup>۱</sup> دانشجوی کارشناسی ارشد، گروه نرم‌افزار، دانشکده کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران،

ymashayekhi@aut.ac.ir

<sup>۲</sup> استاد، گروه نرم‌افزار، دانشکده کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران،

mmeybodi@aut.ac.ir

## چکیده

در سال‌های اخیر شبکه‌های اجتماعی محبوبیت بالایی میان کاربران اینترنت یافته است و به همین دلیل مطالعات زیادی در زمینه تحلیل شبکه‌های اجتماعی انجام شده است. یکی از موضوعات مهم در تحلیل شبکه‌های اجتماعی، بررسی انتشار اطلاعات است. در این زمینه فرض می‌شود که رفتار کاربران تحت تاثیر دیگر کاربران شبکه اجتماعی است. مدل‌های مختلفی برای شبیه‌سازی و تحلیل نحوه انتشار اطلاعات در شبکه‌های اجتماعی طراحی شده است. در این مقاله، ما به بررسی مساله یادگیری احتمالات انتشار برای مدل آبخاری مستقل می‌پردازیم. در ابتدا اهمیت موضوع را بیان می‌کنیم، سپس روش جدیدی برای حل این مساله بر پایه طراحی معادله‌ای برای تخمین احتمال انتشار هر یال، پیشنهاد می‌کنیم. همچنین بهبودی بر روش پیشنهادی به هنگام عدم وجود داده کافی برای یادگیری احتمال انتشار، ارائه می‌گردد. سپس روش پیشنهادی را به کمک چند مجموعه داده ارزیابی کرده و نتایج حاصل از آن را ذکر می‌کنیم. همچنین روش ارائه شده در این مقاله با دیگر روش‌ها از نظر میانگین خطای مطلق و زمان اجرا مقایسه شده است.

## کلمات کلیدی

مدل آبخاری مستقل، انتشار اطلاعات، تحلیل شبکه‌های اجتماعی

گرفته است. در مطالعه انتشار اطلاعات، هدف، کسب دانش در مورد چگونگی انتشار محتوا در شبکه اجتماعی از طریق تعاملات کاربران است.

مطالعه در زمینه انتشار اطلاعات ابتدا در علوم اجتماعی آغاز شد و پس از آن به علوم دیگر نیز گسترش یافت. با رشد شبکه‌های اجتماعی آنلاین در میان مردم، درک رفتار کاربر و نحوه گسترش اطلاعات در شبکه اجتماعی به یک حیطه تحقیقاتی مهم تبدیل شده است. کاربران عمدتاً توسط دوستان خود تحت تاثیر قرار می‌گیرند که نتیجه آن انتشار اطلاعات میان دوستان در شبکه اجتماعی است [۲]. موضوعات اصلی در مطالعه انتشار اطلاعات عبارتند از: شناسایی موضوعات محبوب، شناسایی انتشاردهندگان تاثیرگذار و مدل‌سازی انتشار اطلاعات [۳]. بسیاری از مدل‌های انتشار اطلاعات، تعمیمی بر مدل آبخاری مستقل (IC) [۴] و مدل آستانه خطی (LT) [۵] هستند. انتشار اطلاعات در این مدل‌ها در طی یک فرایند تکراری اتفاق می‌افتد که در

## ۱- مقدمه

در سال‌های اخیر، شبکه‌های اجتماعی آنلاین به یکی از منابع اصلی اطلاعات تبدیل شده است. از جمله این شبکه‌ها فیس‌بوک، گوگل پلاس، توئیتر و غیره می‌باشند. اهمیت فزاینده شبکه‌های اجتماعی، ناشی از حجم زیاد محتوای تولید شده توسط آن‌ها است. کاربران این شبکه‌ها با استفاده از قابلیت‌هایی نظیر به اشتراک گذاری محتوا، بازتوییت، پسندیدن و نظر دادن اطلاعات زیادی تولید می‌کنند، که از این اطلاعات می‌توان به منظور تحلیل رفتار کاربران شبکه اجتماعی استفاده کرد. مسائل مختلفی مانند تشخیص جوامع، پیش‌بینی پیوند و یا انتشار اطلاعات [۱] در این زمینه مورد مطالعه قرار

آن امکان دسترسی یک کاربر به اطلاعات تنها از طریق همسایگانش که در گذشته اطلاعات را دریافت کرده‌اند، وجود دارد [۶-۸]. برخی از مدل‌ها برای در نظر گرفتن زمان پیوسته در فرایند انتشار پیشنهاد شده‌اند [۹ و ۱۰]. این مدل‌ها یک تأخیر زمانی برای انتشار محتوا به یک گره در نظر می‌گیرند. چالش اصلی استفاده از مدل‌های انتشار اطلاعات، یادگیری پارامترهای مدل است. پارامترهای مدل‌های انتشار اطلاعات معمولاً از طریق مشاهده رفتار کاربران شبکه اجتماعی یادگیری می‌شود. پس از یادگیری پارامترها، مدل را می‌توان در برنامه‌های کاربردی مختلف انتشار اطلاعات مانند پیش‌بینی اینکه در انتهای یک فرایند انتشار چه کاربرانی به اطلاعات دسترسی پیدا می‌کنند، پیشینه‌سازی نفوذ در شبکه اجتماعی، جلوگیری از انتشار اطلاعات و غیره به کار برد.

در این مقاله، تمرکز ما بر مدل IC است و سعی می‌کنیم پارامترهای این مدل که در واقع احتمال انتشار اطلاعات یال‌های شبکه اجتماعی است، یاد بگیریم. یادگیری این احتمالات به کمک مشاهده داده‌های انتشار گذشته شبکه انجام می‌شود. مساله یادگیری احتمالات انتشار برای مدل IC در [۱۱] مورد مطالعه قرار گرفته است. ما سعی می‌کنیم احتمالات انتشار را با خطایی کمتر و زمان اجرای بهتر نسبت به روش پایه [۱۱] یاد بگیریم. راه حل ارائه شده در این مقاله یافتن معادله‌ای برای محاسبه احتمالات انتشار یال‌های شبکه اجتماعی است.

قسمت‌های مختلف مقاله به صورتی که در ادامه می‌آید، بخش‌بندی شده است. بخش ۲ کارهای پیشین در مورد این مساله و قدرت و ضعف آنها را بررسی می‌کند. بخش ۳ مساله را بیان و تعریف می‌کند، که شامل توضیح مدل انتشار و تعریف مساله است. بخش ۴ راه حل پیشنهادی را بیان می‌کند. بخش ۵ به ارزیابی روش پیشنهادی برای تخمین احتمالات انتشار می‌پردازد. بخش ۶ شامل نتیجه‌گیری است و پیشنهادهای برای کارهای آینده در این زمینه ارائه می‌دهد.

## ۲- کارهای پیشین

شبکه اجتماعی را می‌توان توسط یک گراف مدل‌سازی کرد، که در آن گره‌های گراف نشان‌دهنده کاربران شبکه اجتماعی هستند و یال‌های گراف روابط بین کاربران مانند دوستی، دنبال کردن و غیره را نشان می‌دهند. در سال‌های اخیر، انتشار اطلاعات در بسیاری از کارهای تحقیقاتی در زمینه تحلیل شبکه‌های اجتماعی مورد مطالعه قرار گرفته است [۶ و ۱۲-۱۴]. در مطالعه انتشار اطلاعات، اغلب، فرایند انتشار توسط مجموعه‌ای از کاربران اولیه آغاز می‌شود و از طریق تعاملات میان کاربران در شبکه گسترش می‌یابد. در اکثر کارها تأثیر خارجی بر شبکه اجتماعی در نظر گرفته نشده است. اما برخی از تحقیقات، تأثیر خارجی را در پخش محتوا در شبکه‌های اجتماعی مورد مطالعه قرار داده‌اند [۱۵].

دمنگوس و ریچاردسون [۱۶ و ۱۷] اولین کسانی بودند که مطالعه انتشار اطلاعات در شبکه‌های اجتماعی و شناسایی گره‌های تأثیرگذار را به کمک روش‌های داده‌کاوی بررسی کردند. کمپه و همکاران [۱۴] به طور رسمی مساله پیشینه‌سازی نفوذ را به عنوان یک مساله بهینه‌سازی تعریف کردند. آن‌ها همچنین اثبات کردند که مساله بهینه‌سازی انتخاب گره‌های تأثیرگذار در مدل IC و مدل LT مساله NP-hard است. آنها یک الگوریتم حریصانه با تضمین ضریب تقریب نیز ارائه دادند.

سایتو و همکاران [۱۱] مساله یادگیری پارامترهای مدل آبخاری مستقل را مورد مطالعه قرار دادند. آن‌ها این مساله را با استفاده از یک تابع درست‌نمایی<sup>۲</sup> بیان کردند و احتمالات انتشار یال‌ها را با استفاده از الگوریتم تکراری پیشینه‌سازی انتظار<sup>۳</sup> (EM) محاسبه کردند. در یکی دیگر از کارهای شاخص یادگیری پارامترهای مدل‌های انتشار، گویال و همکاران [۱۸] یادگیری پارامترهای مدل آستانه عمومی<sup>۴</sup> را مطالعه کردند. بر خلاف روش ارائه شده در [۱۱] روش پیشنهادی [۱۸] مقیاس پذیر است، زیرا از الگوریتم تکراری مانند EM برای یادگیری استفاده نکرده است. بسیاری از کارهای اخیر تلاش کردند نقاط ضعف این دو روش پایه را حل کنند. مدل آبخاری مستقل آسنکرون (AsIC) و مدل آسانه خطی آسنکرون (AsLT) پیشنهاد شده‌اند که زمان را به صورت پیوسته در نظر می‌گیرند. همچنین الگوریتم‌هایی برای یادگیری پارامترهای هر دو مدل طراحی شده است تا امکان مدل‌سازی انتشار اطلاعات در زمان پیوسته فراهم گردد [۹ و ۱۰]. برخی مطالعات مساله یادگیری پارامترهای مدل‌های انتشار اطلاعات را بدون در اختیار داشتن گراف شبکه اجتماعی بررسی کرده‌اند که در این صورت یال‌های گراف شبکه اجتماعی مورد مطالعه، در دسترس نیست [۱۹ و ۲۰].

در دسته دیگری از مطالعات، کارهای متعددی در مورد پویایی انتشار اطلاعات انجام شده است [۲۱-۲۳]. گویل و حسید [۲۳] یک مدل آبخاری مستقل آسنکرون مبتنی بر زمان پیشنهاد داده‌اند که در آن احتمال انتشار وابسته به زمان است. آن‌ها از روش‌های یادگیری ماشین برای یادگیری احتمالات انتشار به کمک ویژگی‌های استخراج شده از شبکه اجتماعی تویتیر استفاده کرده‌اند.

در این مقاله، مساله یادگیری احتمالات انتشار برای مدل IC را مورد بررسی قرار می‌دهیم و روش جدیدی برای حل آن ارائه می‌کنیم. برای تخمین احتمالات انتشار یال‌ها، یک معادله بر حسب داده‌های استخراج شده از انتشارهای گذشته شبکه در نظر می‌گیریم. رویکرد ما به دلیل تخمین افزایشی احتمالات یال‌ها مقیاس‌پذیر است. همچنین نتایج آزمایشات نشان می‌دهد که دقت روش پیشنهاد شده در این مقاله نسبت به روش ارائه شده در [۱۱] بیشتر است.

## ۳- بیان مساله

در این بخش به بیان و تشریح مساله می‌پردازیم. در ابتدا مدل انتشار استفاده شده در این مقاله شرح داده می‌شود. سپس مساله به صورت دقیق تعریف می‌شود.

### ۳-۱- مدل انتشار

در این مقاله، ما تلاش می‌کنیم پارامترهای مدل شناخته شده‌ی آبخاری مستقل را یاد بگیریم. در مدل IC، فرایند انتشار توسط مجموعه‌ای از گره‌های اولیه که تنها گره‌های فعال (گره‌ای که در فرایند انتشار، داده به آن رسیده باشد و یا عملی را که منتشر می‌شود، انجام داده باشد) گراف در مرحله زمانی صفر هستند شروع می‌شود. همچنین فرایند انتشار در مراحل زمانی گسسته رخ می‌دهد. در مرحله زمانی  $t+1$ ، هر گره که در مرحله زمانی  $t$  فعال است، تنها یک شانس برای فعال کردن هر کدام از گره‌های فرزند غیر فعال خود دارد. مستقل از موفقیت و یا شکست گره پدر در فعال نمودن گره فرزند، گره پدر شانس دیگری برای فعال نمودن این گره فرزند ندارد. فرایند انتشار تا زمانی

که گره جدیدی فعال نشود ادامه پیدا می کند. در مدل IC فرض بر این است که یک گره فعال تا انتهای فرآیند انتشار فعال می ماند، به عبارت دیگر، وضعیت گره پس از فعال شدن تغییر نخواهد کرد. تلاش برای فعال نمودن گره فرزند توسط گره پدر با احتمال انتشاری که از قبل مشخص شده است، موفق می شود. بنابراین در مدل IC مشخص کردن احتمال انتشار یال ها ضروری است. در این مقاله، سعی می کنیم این احتمالات را یادگیری نماییم.

### ۳-۲- تعریف مساله

در این مقاله از نشانه گذاری<sup>۵</sup> ارائه شده در [۱۱] استفاده می کنیم. یک شبکه که دارای روابط جهت دار است، توسط گراف  $G = (V, E)$  نمایش داده می شود.  $V$  مجموعه گره های گراف و  $E$  مجموعه یال های گراف را نمایش می دهند. ما هر یال را با  $(u, v)$  نمایش می دهیم که در این حالت که در این حالت یک یال از گره  $u$  به گره  $v$  وجود دارد.

یک نمونه انتشار<sup>۶</sup>  $D = \langle D(0), D(1), \dots, D(T) \rangle$  یک توالی از مجموعه گره ها است، به این صورت که  $D(t)$  گره های فعال گراف در زمان  $t$  را مشخص می کند. فرض بر این است که هر گره می تواند تنها یک بار در یک نمونه انتشار ظاهر شود. به این معنی که هر گره تنها یک بار در یک نمونه انتشار می تواند فعال شود. بنابراین برای یک نمونه انتشار مانند  $D$ ، خواهیم داشت  $D(i) \cap D(j) = \emptyset, i \neq j$ . فرض دیگر این است که هنگامی که یک گره در یک نمونه انتشار فعال می شود، وضعیتش را نمی توان به غیرفعال تغییر داد، بنابراین در آن نمونه انتشار فعال می ماند.

مساله به این صورت است که با در اختیار داشتن گراف  $G = (V, E)$  و یک مجموعه از نمونه های انتشار  $S = \{D_1, D_2, \dots, D_n\}$ ، چگونه این احتمالات را می توان تخمین زد؟ در این مقاله، ما قصد داریم به این سوال پاسخ دهیم و یک روش برای تخمین احتمال انتشار هر یال ارائه کنیم.

### ۴- راه حل پیشنهادی

در این بخش، روش ارائه شده در این مقاله برای یادگیری احتمالات انتشار یال ها در مدل IC توضیح داده می شود.

به منظور یافتن معادله ای جهت تخمین احتمال انتشار هر یال، از نمونه های انتشار به منظور یادگیری استفاده می کنیم. هر نمونه انتشار در مورد احتمال انتشار برخی از یال های گراف اطلاعاتی در اختیار می گذارد.

### ۴-۱- اطلاعات یک نمونه انتشار

در بررسی یک نمونه انتشار، با سه نوع یال مواجه هستیم که اطلاعاتی در مورد احتمال انتشارشان در اختیار می گذارند. یال هایی مانند  $(u, v)$  در یک نمونه انتشار که در آن ها گره  $u$  اصلاً فعال نشده است یا گره  $v$  قبل از گره  $u$  یا همزمان با آن فعال شده است اطلاعاتی در مورد احتمال انتشار این یال در اختیار نمی گذارد. در ادامه به توضیح سه نوع یال در یک نمونه انتشار که اطلاعاتی در مورد احتمال انتشارشان در اختیار می گذارند، می پردازیم.

گونه اول، یال هایی مانند  $(u, v)$  هستند که در آن ها گره  $u$  در مرحله زمانی  $t$  فعال شده است. سپس در مرحله زمانی  $t+1$ ، گره  $v$  فعال شده و پدر دیگری از گره  $v$  در مرحله زمانی  $t$  فعال نشده باشد. پس اطمینان داریم که گره  $u$  در فعال نمودن گره  $v$  موفق بوده است.

نوع دوم یال های گراف در یک نمونه انتشار، شامل یال هایی مانند  $(u, v)$  است که در آن ها گره پدر  $u$  در مرحله زمانی  $t$  فعال شده اما گره فرزند  $v$  بازه زمانی  $0$  تا  $t+1$  فعال نشده باشد. در این حالت می توان نتیجه گرفت که گره  $u$  قطعاً در فعال نمودن گره  $v$  شکست خورده است.

نوع سوم یال ها در یک نمونه انتشار، یال هایی مانند  $(u, v)$  هستند که در آن ها گره پدر  $u$  در مرحله زمانی  $t$  و گره فرزند  $v$  در مرحله زمانی  $t+1$  فعال شده اند. همچنین پدرهای دیگری از  $v$  نیز در مرحله زمانی  $t$  فعال شده اند. در این حالت مشخص نیست که کدام یک از پدرهای  $v$  این گره را فعال نموده اند، اما می دانیم یکی از پدرها در فعال کردن گره  $v$  موفق شده است و بقیه یا در فعال کردن گره  $v$  شکست خورده اند و یا قبل از آن که شانس فعال کردن این گره را داشته باشند، گره  $v$  توسط پدر دیگری فعال شده است.

### ۴-۲- تخمین احتمال انتشار

احتمال انتشار در یک یال مانند  $(u, v)$  با مشاهده یک مجموعه نمونه های انتشار یادگیری، برابر است با تعداد دفعاتی که گره  $u$  موفق شده است گره  $v$  را فعال کند، تقسیم بر تعداد دفعات تلاش گره  $u$  برای فعال نمودن گره  $v$ . مشکل در استفاده از این معادله زمانی است که مطمئن نیستیم کدام یک از پدرهای فعال شده در مرحله قبل گره  $v$  آن را فعال کرده اند. در واقع موفقیت و یا شکست پدرها را در فعال نمودن آن نمی دانیم. همچنین رخ دادن تلاش برای فعال کردن گره  $v$  را نیز از سوی هر کدام از این پدرها نمی دانیم.

در این مقاله فرض می کنیم که تمام پدرهای فعال شده در مرحله قبل گره  $v$  آن را فعال نموده اند، با این تفاوت که به اطلاعات به دست آمده از این نمونه انتشار وزن کمتری نسبت به نمونه های انتشاری که مطمئن هستیم گره پدر  $u$  قطعاً در فعال نمودن گره فرزند  $v$  موفق شده است یا شکست خورده است، می دهیم. به اطلاعات نمونه های انتشاری که مطمئن هستیم گره پدر  $u$  قطعاً در فعال نمودن گره فرزند  $v$  موفق شده است یا شکست خورده است، وزن یک می دهیم، و به اطلاعات نمونه هایی که گره فرزند  $v$  در یک نمونه انتشار مانند  $D_s$  در زمان  $t+1$  فعال شده است و تعداد  $k_s(v)$  پدر آن در زمان  $t$  فعال شده اند  $(k_s(v) > 1)$ ، وزن  $\frac{1}{k_s(v)}$  می دهیم.

فرض  $C_s(t)$  نشان دهنده گره هایی از گراف باشد که در نمونه انتشار  $D_s$  قبل از زمان  $t+1$  فعال شده اند. همچنین  $AT_s(u)$  زمان فعال شدن گره  $u$  در نمونه انتشار  $D_s$  را مشخص کند. در صورتی که یک گره در نمونه انتشار  $D_s$  فعال نشده باشد، این مقدار برای آن گره بی نهایت است. همچنین به طور دقیق تر داریم  $k_s(v) = |\{u | (u, v) \in E, AT_s(u) = AT_s(v) - 1\}|$ .  $T(D_s)$  آخرین زمان فعالیت گره های نمونه انتشار  $D_s$  را نشان می دهد. در نتیجه معادله پیشنهادی برای تخمین احتمال انتشار یک یال مانند  $(u, v)$  که با  $p_{uv}$  نمایش داده می شود، با در اختیار داشتن یک مجموعه نمونه های انتشار مانند  $S = \{D_1, D_2, \dots, D_n\}$ ، در معادله (۱) آمده است:

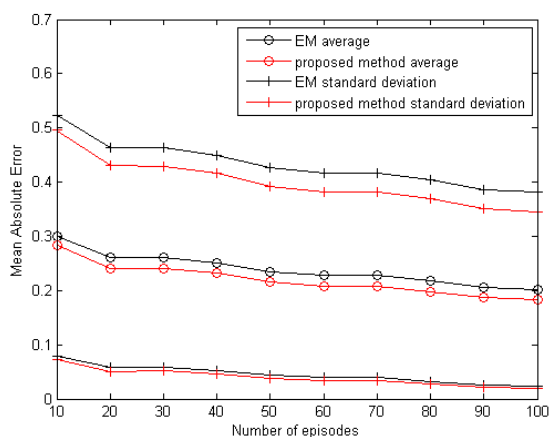
$$p_{uv} = \frac{\sum_{D_s \in S, AT_s(u) \leq T(D_s), AT_s(v) = AT_s(u)+1} 1/k_s(v)}{\left( \sum_{D_s \in S, AT_s(u) \leq T(D_s), AT_s(v) > AT_s(u)+1} 1 + \sum_{D_s \in S, AT_s(u) \leq T(D_s), AT_s(v) = AT_s(u)+1} 1/k_s(v) \right)} \quad (1)$$

به شکل مشابه [۱۱]، از بازه [۰،۱،۰،۳] به صورت تصادفی یکنواخت انتخاب شد.

روش پیشنهادی در این مقاله را برای تخمین احتمالات انتشار یال‌های گراف با الگوریتم پیشنهاد شده در [۱۱] مقایسه کردیم. سایتو و همکاران [۱۱] یک روش برای تخمین احتمالات انتشار با مشاهده مجموعه‌ای از نمونه‌های انتشار طراحی کرده‌اند. در این روش یک تابع درست‌نمایی طراحی شده و برای بهینه‌سازی آن از الگوریتم EM استفاده شده است. ما روش‌ها را با توجه به میانگین خطای مطلق<sup>۴</sup> (MAE) احتمالات تخمین زده شده نسبت به احتمالات اصلی یال‌ها مقایسه می‌کنیم. همچنین روش‌ها از نظر زمان اجرا نیز مورد ارزیابی قرار می‌گیرند.

## ۲-۵- نتایج

در این بخش، آزمایشات و نتایج حاصل از آن‌ها ارائه می‌گردد. همچنین نتایج روش پیشنهادی با روش [۱۱] که این مساله را حل می‌کند، مقایسه می‌شود. شکل (۱)، (۲)، (۳) و (۴) نمودار MAE (بر اساس اختلاف احتمال انتشار تخمین زده شده برای هر یال با احتمال انتشار اصلی یال) و انحراف معیار خطا بر حسب تعداد نمونه‌های انتشار یادگیری برای هر یک از مجموعه داده‌ها را نشان می‌دهند. همانطور که مشخص است، روش پیشنهادی خطای کمتری نسبت به روش [۱۱] دارد و در نتیجه دقت بالاتری در تخمین احتمالات انتشار یال‌های گراف با مشاهده مجموعه نمونه‌های انتشار دارد. همچنین، مشاهده می‌کنیم که با افزایش تعداد نمونه‌های انتشار یادگیری، خطای روش‌ها کمتر می‌شود که نشان از اهمیت میزان داده یادگیری دارد.



شکل (۱): نمودار MAE بر حسب تعداد نمونه‌های انتشار مجموعه داده Digg

همچنین حالتی که گره  $v$  در مرحله زمانی  $t+1$  فعال شود و تنها یک پدر مانند  $u$  دارد که در مرحله زمانی  $t$  فعال شده باشد نیز پوشش داده می‌شوند، زیرا وزن  $\frac{1}{k_s(v)}$  برابر با ۱ می‌شود که در صورت و مخرج کسر احتمال، ظاهر شده است.

## ۳-۴- بهبود روش پیشنهادی

یکی از مشکلات تخمین احتمال انتشار برای یال‌های گراف مانند یال  $(u, v)$ ، هنگامی است که در هیچ یک از نمونه‌های انتشار یادگیری، حالتی که مطمئن باشیم گره  $u$  در فعال نمودن گره  $v$  موفق شده باشد یا شکست خورده باشد، وجود ندارد. در این حالت تنها اطلاعات در دسترس حالتی است که گره  $v$  در مرحله زمانی  $t+1$  فعال شده است و گره  $u$  در مرحله زمانی  $t$ ، اما پدران دیگری از  $v$  نیز در مرحله زمانی  $t$  فعال شده‌اند، در نتیجه مطمئن نیستیم آیا گره  $u$  گره  $v$  را فعال کرده است یا خیر. اگر برای یک یال تنها چنین داده‌هایی در دسترس باشد، احتمال انتشار یال را از روی داده‌ها تخمین نمی‌زنیم، زیرا چنین داده‌هایی داده‌های محکم و مطمئنی برای تخمین احتمال انتشار یک یال نیستند، در نتیجه یک مقدار تصادفی یکنواخت به عنوان احتمال انتشار یال در نظر می‌گیریم.

## ۵- ارزیابی

در این بخش، آزمایشات و نتایج حاصل از آن‌ها ارائه می‌گردد. همچنین نتایج روش پیشنهادی با روش پایه [۱۱] که همین مساله را حل می‌کند، مقایسه می‌شود.

## ۵-۱- شرایط آزمایش

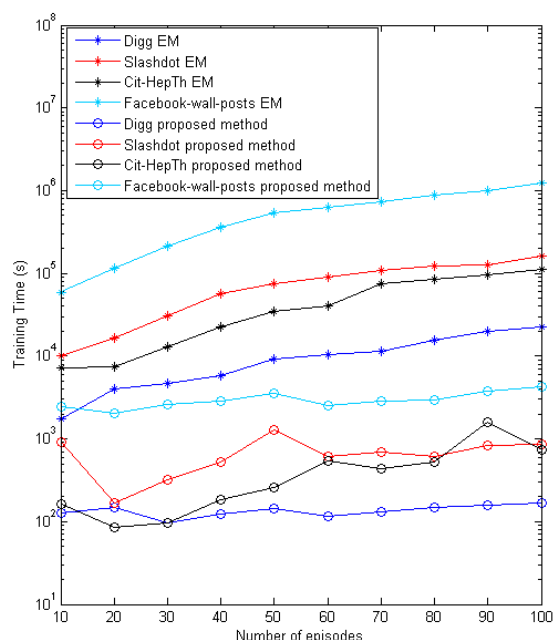
جدول (۱) مجموعه داده‌های مورد استفاده جهت ارزیابی را نشان می‌دهد. یال‌های طوقه<sup>۵</sup> (یال‌هایی که رئوس ابتدا و انتهای آن یکسان هستند) از این شبکه‌ها حذف شدند. مجموعه داده‌های مورد استفاده شامل Digg [۲۴]، Slashdot [۲۵]، Cit-HepTh [۲۶] و Facebook-wall-posts [۲۷] هستند.

جدول (۱): مجموعه داده‌ها

نام مجموعه داده	تعداد یال	تعداد گره	توضیح
Digg	۸۶۴۰۴	۳۰۳۹۸	شبکه پاسخ مربوط به وبسایت اخبار اجتماعی Digg
Slashdot	۱۳۱۱۷۵	۵۱۰۸۳	شبکه پاسخ مربوط به وبسایت فناوری ارتباطات Slashdot
Cit-HepTh	۳۵۲۸۰۷	۲۷۷۷۰	شبکه مقالات در arXiv's High Energy Physics
Facebook-wall-posts	۲۷۴۰۸۶	۴۶۹۵۲	شبکه پست‌های مجموعه کوچکی از کاربران فیسبوک در صفحه یکدیگر

برای هر مجموعه داده، ۱۰۰ نمونه انتشار با شروع از یک گره انتخاب شده به صورت تصادفی یکنواخت، تولید شد. همچنین روش‌ها را با تعداد متفاوت نمونه انتشارهای یادگیری، مقایسه می‌کنیم. احتمالات انتشار هر یال

استفاده می‌کند و تا زمان همگرایی چندین بار برای تمامی یال‌ها محاسبه احتمالات انتشار یال‌ها مجدد تکرار می‌شود. در صورتی که در روش پیشنهادی این مقاله، الگوریتم برای هر نمونه انتشار و هر یال که در آن نمونه انتشار، اطلاعاتی در موردش می‌توان به دست آورد، یک بار عمل به‌روزرسانی احتمالات را انجام می‌دهد و مجدد به این نمونه انتشار برای این یال بر نمی‌گردد.

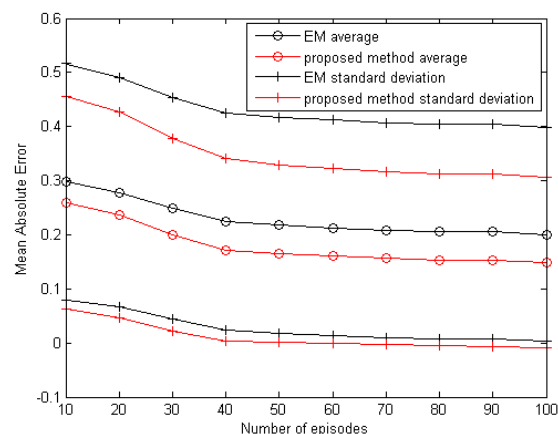


شکل (۵): نمودار زمان اجرا (مقیاس لگاریتمی) بر حسب تعداد نمونه‌های انتشار برای تمام مجموعه داده‌ها

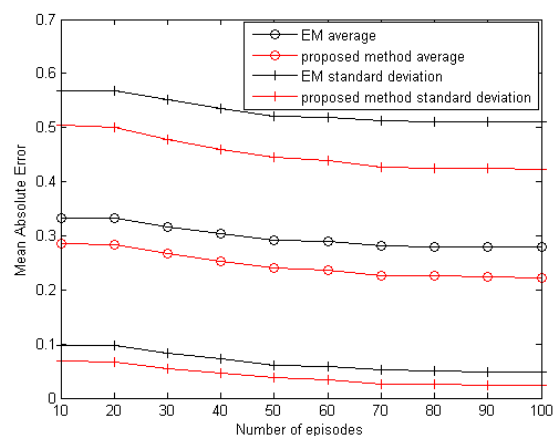
## ۶- جمع‌بندی و کارهای آتی

در سال‌های اخیر مطالعات فراوانی در زمینه تحلیل شبکه‌های اجتماعی صورت گرفته است. یکی از موضوعات در این زمینه، بررسی انتشار اطلاعات است. در این مقاله روش جدیدی برای یادگیری احتمالات انتشار در مدل آشنایی مستقل ارائه گردید. در این روش برای محاسبه احتمال هر یال، یک معادله بر اساس مشاهده یک مجموعه از نمونه‌های انتشار پیشنهاد شده است. همچنین بهبودی بر روش پیشنهادی به منظور جلوگیری از تخمین اشتباه احتمالات انتشار تحت تأثیر مجموعه داده آموزشی نامتوازن، ارائه شد. با آزمایشات انجام شده می‌توان نتیجه گرفت روش پیشنهادی در این مقاله احتمالات انتشار برای مدل آشنایی مستقل را با دقت بیشتر و زمان اجرای بسیار کم‌تری نسبت به روش [۱۱] یادگیری می‌کند.

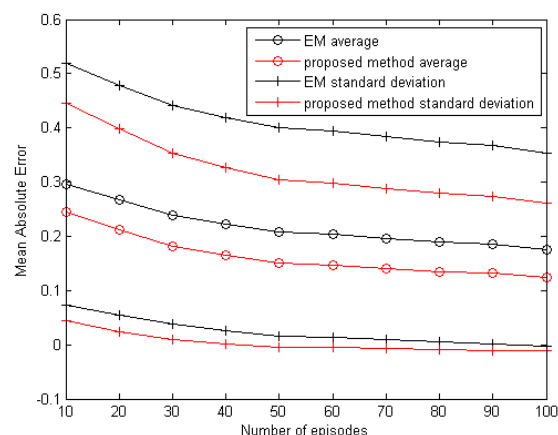
نتایج مطلوب روش پیشنهادی این مقاله باعث می‌شود که بتوانیم امیدوار به تعمیم‌های مختلفی بر این روش باشیم. به عنوان مثال هم‌اکنون در حال بهبود این روش در شرایطی که نتوان اطلاعات زیادی را برای یادگیری برخی از احتمالات انتشار به دست آورد، هستیم. همچنین سعی داریم روشی مشابه برای در نظر گرفتن زمان پیوسته در مدل انتشار، طراحی کنیم.



شکل (۲): نمودار MAE بر حسب تعداد نمونه‌های انتشار مجموعه داده Slashdot



شکل (۳): نمودار MAE بر حسب تعداد نمونه‌های انتشار مجموعه داده Cit-HepTh



شکل (۴): نمودار MAE بر حسب تعداد نمونه‌های انتشار مجموعه داده Facebook-wall-posts

شکل (۵) زمان اجرا بر حسب تعداد نمونه‌های انتشار یادگیری روش‌ها برای تمام مجموعه داده‌ها را نشان می‌دهد. همانطور که در نمودارها مشخص است، زمان اجرای روش پیشنهادی بسیار کم‌تر از روش [۱۱] است. دلیل این مطلب این است که روش [۱۱] از الگوریتم EM که یک روش تکراری است

- SIGKDD international conference on Knowledge discovery and data mining*, 2001, pp. 57-66.
- [17] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing." in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 61-70.
- [18] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks." in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 241-250.
- [19] S. Lamprier, S. Bourigault, and P. Gallinari, "Influence learning for cascade diffusion models: focus on partial orders of infections." *Social Network Analysis and Mining*, vol. 6, p. 93, 2016.
- [20] S. Bourigault, S. Lamprier, and P. Gallinari, "Representation learning for information diffusion through social networks: an embedded cascade model." in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 2016, pp. 573-582.
- [21] D. Li, S. Zhang, X. Sun, H. Zhou, S. Li, and X. Li, "Modeling information diffusion over social networks for temporal dynamic prediction." *IEEE Transactions on Knowledge and Data Engineering*, 2017.
- [22] K. Zhuang, F. Han, H. Shen, K. Zhang, and H. Zhang, "Information Diffusion Temporal Dynamic Prediction in Microblog System Based On User Influence Learning." *International Journal of Hybrid Information Technology*, vol. 9, pp. 327-336, 2016.
- [23] A. Guille and H. Hacid, "A predictive model for the temporal dynamics of information diffusion in online social networks." in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 1145-1152.
- [24] M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann, "Social synchrony: Predicting mimicry of user actions in online social media." in *Computational Science and Engineering, 2009. CSE'09. International Conference on*, 2009, pp. 151-158.
- [25] V. Gómez, A. Kaltenbrunner, and V. López, "Statistical analysis of the social network and discussion threads in Slashdot." in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 645-654.
- [26] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, p. 2, 2007.
- [27] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook," in *Proceedings of the 2nd ACM workshop on Online social networks*, 2009, pp. 37-42.
- [1] C. Lagnier, L. Denoyer, E. Gaussier, and P. Gallinari, "Predicting information diffusion in social networks using content and user's profiles." in *European conference on information retrieval*, 2013, pp. 74-85.
- [2] Q. Bao, W. K. Cheung, Y. Zhang, and J. Liu, "A component-based diffusion model with structural diversity for social networks." *IEEE transactions on cybernetics*, vol. 47, pp. 1078-1089, 2017.
- [3] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey." *ACM SIGMOD Record*, vol. 42, pp. 17-28, 2013.
- [4] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth." *Marketing letters*, vol. 12, pp. 211-223, 2001.
- [5] M. Granovetter, "Threshold models of collective behavior." *American journal of sociology*, vol. 83, pp. 1420-1443, 1978.
- [6] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks." in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, 2010, pp. 599-608.
- [7] M. Gomez Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks." in *Proceedings of the 28th international conference on machine learning (ICML- 11), ICML'11, ACM*, 2011, pp. 561-568.
- [8] G. Ver Steeg and A. Galstyan, "Information-theoretic measures of influence based on content dynamics." in *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 3-12.
- [9] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Learning continuous-time information diffusion model for social behavioral data analysis." *Advances in Machine Learning*, pp. 322-337, 2009.
- [10] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Selecting information diffusion models over social networks for behavioral analysis." in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2010, pp. 180-195.
- [11] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model." in *Knowledge-based intelligent information and engineering systems*, 2008, pp. 67-75.
- [12] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence." *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, p. 21, 2012.
- [13] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace." in *Proceedings of the 13th international conference on World Wide Web*, 2004, pp. 491-501.
- [14] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network." in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 137-146.
- [15] S. A. Myers, C. Zhu, and J. Leskovec, "Information diffusion and external influence in networks." in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 33-41.
- [16] P. Domingos and M. Richardson, "Mining the network value of customers." in *Proceedings of the seventh ACM*

## پانویس ها

- 
- <sup>1</sup> Information diffusion  
<sup>2</sup> Likelihood  
<sup>3</sup> Expectation maximization  
<sup>4</sup> General threshold model  
<sup>5</sup> Notation  
<sup>6</sup> Diffusion episode  
<sup>7</sup> Self-edge  
<sup>8</sup> Mean absolute error