

به نام خدا

مساله شناسایی کاربران مهم در شبکه های اجتماعی

مهسا دیباجی - زهرا نکونام

## مقدمه :

مساله ماکسیم سازی تاثیر<sup>1</sup> به مشخص کردن موثرترین اشخاص در یک شبکه اجتماعی میپردازد و در موضوعاتی مثل امنیت شبکه، تشخیص ویروس کامپیوتر و غیره هم از آن استفاده می شود.

IMP را میتوان به صورت  $\max\{\sigma(S) : |S|=k, S \subset V\}$  روی یک شبکه تصادفی در نظر گرفت. که  $V$  مجموعه رئوس را نشان می دهد و  $S$  زیرمجموعه ای از  $V$  است که به آن مجموعه بذر (seed set) می گوییم و اندازه آن  $k$  است. تابع هدف این مساله نیز  $\sigma(S)$  است که در ادامه به توضیح آن خواهیم پرداخت.

## مفروضات مسئله:

در هر مساله IMP دو بخش مهم وجود دارد. با داشتن شبکه  $N$  و یک مدل انتشار، IMP به دنبال شناسایی مجموعه بذر (seed set) بهینه ای هستیم که اندازه آن  $k$  باشد و هنگامی که فرآیند انتشار با نودهای مجموعه  $S$  آغاز می شود، تابع تاثیر  $\sigma(S)$  ماکسیمم شود.

IMP روی یک شبکه وزن دار و جهت دار  $N=(V, A, W)$  تعریف می شود.

- $V$ : مجموعه نودهایی است که نشان دهنده اعضای شبکه اجتماعی هستند.  $|V|=n$
- $A$ : روابط در این شبکه مانند دنبال کردن (follow)، دوستی در شبکه های اجتماعی یا روابط مشابه با کمان ها نمایش داده می شوند که  $A$  مجموعه این کمان هاست.  $|A|=m$
- $W$ : مجموعه ی شامل وزن های متناظر با کمان هاست. هر کمان  $(i,j) \in A$  دارای وزن  $P_{ij}$  است که نشان دهنده احتمال این که نود  $i$ ، نود  $j$  را هنگامی که یک action از سوی نود  $i$  آغاز میشود تحت تاثیر قرار دهد؛ می باشد.
- **مجموعه بذر<sup>2</sup>  $S$** : مجموعه ی نود های فعال (موثر) در شروع فرآیند آشنایی.

<sup>1</sup> influence maximization problem

<sup>2</sup> Seed set

- $\sigma(S)$ : اگر یک مجموعه  $S$  به عنوان مجموعه بذر داشته باشیم؛ تعداد مورد انتظار نودهایی که در انتهای فرآیند آبخاری<sup>۳</sup> فعال شده اند/تحت تاثیر قرار گرفته‌اند توسط تابع  $\sigma(S)$  محاسبه می شود.

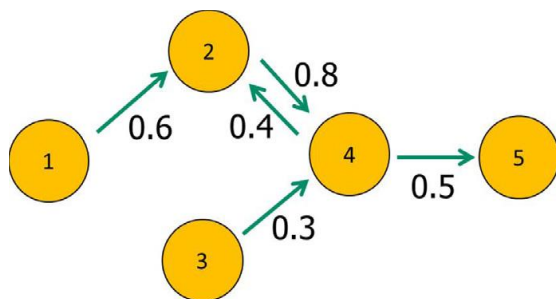
**مدل انتشار:** فرض می شود اطلاعات در شبکه تحت تاثیر قوانین مشخصی پخش می شوند و مدل انتشار نحوه انتشار اطلاعات در شبکه را توصیف می کند. چند مدل شناخته شده وجود دارند که در این گزارش به توضیح مدل انتشار آبخاری مستقل خواهیم پرداخت.

**نود فعال (تحت تاثیر):** صرف نظر از مدل انتشار، یک نود  $i$  را فعال یا تحت تاثیر می نامیم اگر نسبت به اطلاعاتی که با او منتشر شده، علاقه نشان دهد. هنگامی که یک نود فعال می شود تا انتهای فرآیند انتشار فعال می ماند و نمی تواند خود را غیرفعال کند.

### قوانین مدل انتشار آبخاری مستقل<sup>۴</sup> (IC):

- یک نود  $i$  که در مرحله  $t$  فعال شده تنها یک شانس دارد تا هر همسایه  $j$  خود را با احتمال  $P_{ij}$  تحت تاثیر قرار دهد. این تلاش ها از هم مستقل اند و اگر نود  $i$  موفق نشود، دیگر نمی تواند نود  $j$  را تحت تاثیر قرار دهد.
- هنگامی که نود  $i$  در مرحله  $t$  نود  $j$  را فعال می کند، نود  $j$  در مرحله  $t+1$  تلاش می کند همه همسایه های غیرفعالش را فعال کند.
- اگر بیشتر از یک نود به دنبال فعال کردن نود  $j$  بودند، آن ها به طور دلخواه مرتب میشوند.
- فرآیند با مجموعه بذر فعال و اولیه  $S$  آغاز می شود و تازمانی ادامه دارد که هیچ فعال سازی ای ممکن نباشد.

مثال: یک نمونه شبکه با ۵ نود و ۵ کمان را مشاهده میکنید.



<sup>3</sup> Cascading process

<sup>4</sup> Independent Cascade diffusion model

در شکل بالا، اگر فرض کنیم نود ۱ به عنوان مجموعه بذر  $S$  انتخاب شده باشد، این نود با احتمال ۰,۶ می تواند روی تنها همسایه اش، نود ۲ تاثیر بگذارد. اگر موفق شد، در مرحله بعدی نود ۲ با احتمال ۰,۸ نود ۴ را تحت تاثیر قرار می دهد و به همین ترتیب ادامه می یابد؛ به عنوان مثال اگر نود ۴ فعال شود با احتمال ۰,۵ نود ۵ را تحت تاثیر قرار می دهد.

## ارزیابی تأثیر

ثابت می شود از نظر محاسباتی، محاسبه  $\sigma(S)$  به طور دقیق،  $\#P\text{-hard}$  است. اما همانطور که گفته شد  $\sigma(S)$  معادل است با تعداد مورد انتظار (تعداد متوسط) نودهایی که در انتهای فرآیند تحت تاثیر قرار گرفته اند و با دید دیگر می توان گفت تعداد مورد انتظار نود هایی که توسط مجموعه  $S$  یک گراف رندوم، قایل دسترسی هستند و به آن ها مسیر وجود دارد.

برای محاسبه  $\sigma(S)$  همه ی سناریو های ممکن از شبکه تصادفی لحاظ می شوند.

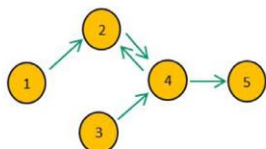
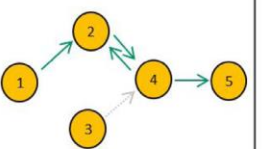
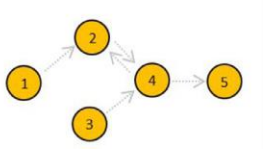
**سناریو:** هر سناریو با یک زیرمجموعه از یال های فعال و غیرفعال در شبکه متناظر است. این که یال  $(i, j)$  در شبکه فعال هست یا نه سناریو های مختلف را نشان می دهد. با توجه به آن که در شبکه  $m$  یال داریم و برای هر یال دو وضعیت وجود دارد پس به طور کلی  $2^m$  سناریو موجود می باشد. چون تعداد یال ها متناهی است در نتیجه تعداد سناریوهای ممکن هم هر چند زیاد اما متناهی است.

**یال فعال:** یک یال را فعال می گوییم اگر در شبکه متناظر با یک سناریو داده شده، موجود باشد. در مدل IC یک یال یا با احتمال  $P_{ij}$  فعال است یا با احتمال  $1 - P_{ij}$  غیرفعال.

مجموعه همه سناریوها را با  $R$  نشان می دهیم و  $r$  را اندیس سناریو داده شده در نظر می گیریم. احتمال رخداد سناریو  $r$  ام را با  $\mu_r$  نشان می دهیم که از رابطه (۱) به دست می آید.

$$\mu_r = \prod_{(i, j) \in \text{activeArcs}} P_{ij} \prod_{(i, j) \notin \text{activeArcs}} (1 - P_{ij}) \quad (1)$$

برای درک بهتر محاسبه  $\sigma(S)$  به توضیح مثال زیر می‌پردازیم:

	$r=1$  $\mu_1=(0.6)(0.4)(0.8)(0.3)(0.5)=0.0288$		$r=2$  $\mu_2=(0.6)(0.4)(0.8)(1-0.3)(0.5)=0.0672$		...	$r=32$  $\mu_{32}=(1-0.6)(1-0.4)(1-0.8)(1-0.3)(1-0.5)=0.0168$		
Seed Set	Influence Spread	Contribution to $\sigma(s)$	Influence Spread	Contribution to $\sigma(s)$	...	Influence Spread	Contribution to $\sigma(s)$	$\sigma(s)$
{1,2}	4	0.1152	4	0.2688	...	2	0.0334	2.7492
{1,3}	5	0.1440	5	0.3360	...	2	0.0334	3.6184
...	...	...	...	...	...	...	...	...
{4,5}	3	0.0864	3	0.2016	...	2	0.0334	2.4086

در این مثال ۵ یال و در نتیجه  $2^5=32$  سناریو داریم که ۳ حالت ممکن از آنها نشان داده شده است.

در سناریوی اول ( $r=1$ ) فرض میکنیم همه یال‌ها فعال هستند و احتمال رخداد این سناریو نیز با توجه به فرمول (۱) به دست می‌آید. ( $\mu_r = 0.0288$ ). در سناریوی دوم ( $r=2$ ) همه یال‌ها به جز یال (3,4) فعال هستند و در سناریوی آخر ( $r=32$ ) هیچ یالی فعال نیست.

فرض کنیم سائز مجموعه بذرمان  $k=2$  باشد. در نتیجه  $\binom{5}{2}=10$  حالت برای انتخاب این مجموعه وجود دارد. همچنین برای هر مجموعه راس و هر سناریو باید گسترش تاثیر<sup>۵</sup> را محاسبه کنیم. برای مثال در سناریو اول هنگامی که انتشار از رئوس {۱,۲} آغاز شود، ۴ راس {۱,۲,۴,۵} فعال خواهند شد. این مقدار با شمارش همه نودهایی که از مجموعه بذر توسط یال‌های فعال قابل دسترسی هستند، محاسبه خواهد شد.

برای هر مجموعه بذر، سهم هر سناریو در  $\sigma(S)$  نیز برابر با حاصل ضرب احتمال رخداد سناریو در تعداد نودهایی است که توسط این مجموعه بذر و یال‌های فعال، فعال شده‌اند.

حال  $\sigma(S)$  برای هر مجموعه بذر  $S$  برابر است با مجموع سهم همه سناریوهای ممکن در  $\sigma(S)$ . در نهایت آن مجموعه بذری که بیشترین  $\sigma(S)$  را تولید میکند به عنوان جواب بهین تعیین می‌شود.

<sup>5</sup> Influence spread

## فرمول ریاضی IMP :

IMP را می توان به صورت  $\{\max \sigma(S) : |S| = k; S \subset V\}$  نوشت اما این روش برای بهینه سازی مناسب نیست چون NP-hard می باشد. همانطور که مشاهده کردیم تعداد سناریوها با افزایش تعداد یالها به صورت نمایی زیاد می شود و حجم محاسبات برای محاسبه دقیق  $\sigma(S)$  زیاد است. در نتیجه مشخص کردن جواب بهین حتی برای شبکه های کوچک هم بسیار سخت است.

### • مجموعه ها :

$V$	مجموعه رئوس $ V  = n$
$A$	مجموعه کمان ها $ A  = m$
$W$	مجموعه وزن های یال ها
$S$	مجموعه نودهای موثر اولیه (مجموعه بذر)
$R$	مجموعه همه ی سناریوهای ممکن
$P_{ir}$	مجموعه همه predecessor های نود $i$ در سناریوی $r$ (predecessor) ها نودهایی هستند که می توانند نود $i$ را فعال کنند).

$P_{ir}$  با انجام یک BFS با شروع از نود  $i$  و حرکت کردن در خلاف جهت با استفاده از کمان های فعال بدست می آید.

### • پارامترها:

$k$	سایز مجموعه بذر
$P_{ij}$	احتمال آن که نود $i$ نود $j$ را تحت تاثیر قرار دهد
$\mu_r$	احتمال رخداد سناریو $r$ ام
$R$	تعداد سناریو های ممکن $ R  = R$

### • متغیرهای تصمیم:

$y_i$	متغیر باینری که اگر نود $i$ به عنوان seed انتخاب شود، یک و در غیر این صورت صفر است.
$x_{ir}$	متغیر باینری که اگر نود $i$ در سناریوی $r$ فعال شود، یک و در غیر این صورت صفر است.

$\chi(y)$ : مجموعه ای از متغیرهای تصادفی متناظر با همه ی نودهایی که در فرآیند انتشار فعال می شوند در صورتی که  $y$  مجموعه بذر (seed set) باشد. در حقیقت اعضای از  $\chi(y)$  که دارای مقدار یک هستند، نود های مجموعه بذر و نودهایی هستند که در مراحل بعدی در فرایند انتشار IC فعال شده اند. در نتیجه توسط این تعاریف  $\sigma(S)$  را به صورت زیر می توان بازنویسی کرد.

$$\sigma(S) = \sigma(y) = \sum_{i \in r} E[x(y)] = \sum_{r \in R} \sum_{i \in V} \mu_r x_{ir}$$

مدل IMBIP<sup>6</sup>:

$$\begin{aligned} \max z &= \sum_{r \in R} \sum_{i \in V} \mu_r x_{ir} \\ s.t. & \\ \sum_{i \in V} y_i &= k \\ x_{ir} &\leq \sum_{j \in p_{ir}} y_j \quad i \in V, r \in R \\ 0 \leq x_{ir} &\leq 1, \quad y_i \in \{0,1\} \quad i \in V, r \in R \end{aligned}$$

قید اول تضمین می کند که سایز مجموعه بذر انتخاب شده  $k$  باشد.

قید دوم تضمین می کند در سناریوی  $r$ ، نود  $i$  در صورتی می تواند فعال شود که یا predecessor ای در مجموعه seed داشته باشد (یعنی توسط کمان های فعالی به یک نود در مجموعه بذر متصل باشد) یا خود در مجموعه seed باشد.

<sup>6</sup> Influence Maximization Binary Integer Program