

Perbandingan Sepuluh Metrik Kemiripan Berbasis Teori Graf dan Operasi Himpunan untuk Rekomendasi Teman pada Jejaring Sosial

Nabil Rabbani - 13525004

Program Studi Teknik Informatika

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung, Jalan Ganesha 10 Bandung

E-mail: takkaloppaeh@gmail.com , 13525004@std.stei.itb.ac.id

Abstrak—Fitur rekomendasi teman pada jejaring sosial, yang di Facebook dikenal sebagai “Orang yang Mungkin Anda Kenal”, pada dasarnya adalah persoalan prediksi tautan, yaitu menebak sisi yang belum muncul pada graf pertemanan. Makalah ini memodelkan jejaring sosial sebagai graf tak-berarah dan menurunkan sepuluh metrik kemiripan pasangan simpul semata-mata dari operasi himpunan pada himpunan tetangga, lalu mengelompokkannya menjadi tiga keluarga: berbasis cacah, ternormalisasi, dan berbobot derajat. Kesepuluh metrik dibandingkan pada data pertemanan Facebook nyata (SNAP ego-Facebook, 4.039 simpul dan 88.234 sisi) menggunakan validasi silang Monte Carlo sebanyak 20 ulangan pemisahan acak dengan data uji 20%, memakai AUC dan presisi pada rekomendasi teratas sebagai ukuran. Indeks Resource Allocation memberikan AUC tertinggi (0,972), diikuti Adamic-Adar (0,962) dan Common Neighbors (0,952); urutan peringkat ini konsisten di seluruh 20 ulangan. Sebagian besar metrik ternormalisasi mencapai AUC tinggi tetapi presisinya buruk pada sedikit rekomendasi teratas. Sebagai temuan tambahan, koefisien Jaccard dan Sorensen terbukti ekuivalen-peringkat secara matematis sehingga menghasilkan AUC identik pada seluruh ulangan.

Kata kunci—graf; himpunan; prediksi tautan; rekomendasi teman; Resource Allocation; validasi silang Monte Carlo

I. PENDAHULUAN

Jejaring sosial daring seperti Facebook, Instagram, dan LinkedIn telah menjadi bagian keseharian banyak orang. Salah satu fitur yang hampir selalu hadir adalah rekomendasi teman, yang di Facebook dikenal sebagai “People You May Know” atau “Orang yang Mungkin Anda Kenal”. Fitur ini menyodorkan daftar pengguna yang kemungkinan besar dikenal seseorang di dunia nyata meskipun keduanya belum terhubung di dalam platform. Kualitas rekomendasi ini berdampak langsung pada pertumbuhan jaringan dan kenyamanan pengguna, sehingga cara menghitungnya menjadi persoalan yang menarik.

Intuisi yang mendasari fitur tersebut sederhana: jika dua orang memiliki banyak teman yang sama, besar kemungkinan mereka juga saling mengenal. Intuisi ini dapat dirumuskan secara tepat menggunakan dua pokok bahasan Matematika Diskrit, yaitu teori graf dan teori himpunan. Pengguna direpresentasikan sebagai simpul, pertemanan sebagai sisi, dan himpunan teman seseorang sebagai himpunan tetangga simpul

tersebut. Pertanyaan “siapa yang mungkin berteman dengan siapa” kemudian berubah menjadi persoalan menebak sisi yang belum muncul pada graf, yang dalam literatur disebut prediksi tautan (link prediction).

Terdapat banyak cara mengubah informasi teman bersama menjadi sebuah skor kemiripan, dan cara-cara tersebut tidak selalu menghasilkan urutan rekomendasi yang sama. Makalah ini berangkat dari pertanyaan praktis: di antara seluruh rumus skor yang dapat dibangun hanya dari himpunan tetangga, manakah yang paling akurat ketika diterapkan pada data pertemanan nyata? Untuk menjawabnya secara menyeluruh, penulis tidak memilih beberapa rumus secara sembarang, melainkan meninjau keluarga lengkap sepuluh metrik yang seluruhnya dapat diturunkan dari operasi himpunan pada himpunan tetangga, yaitu Common Neighbors, koefisien Jaccard, indeks Salton, indeks Sørensen, Hub Promoted Index, Hub Depressed Index, indeks Leicht-Holme-Newman, Resource Allocation, Adamic-Adar, dan Preferential Attachment.

Perlu ditegaskan sejak awal bahwa membandingkan metrik kemiripan untuk prediksi tautan bukan hal baru dalam riset jaringan kompleks [3], [7], [8]. Nilai makalah ini terletak pada penerapan konsep Matematika Diskrit secara mandiri dan dapat diverifikasi, sekaligus pada penambahan kerangka pengujian statistik, yakni validasi silang Monte Carlo, uji tanda, dan analisis sensitivitas, yang jarang muncul pada makalah sejenis di tingkat mata kuliah.

Kontribusi makalah ini ada empat. Pertama, menurunkan sepuluh metrik kemiripan pasangan simpul murni dari operasi himpunan pada himpunan tetangga dan mengelompokkannya secara sistematis ke dalam tiga keluarga. Kedua, membandingkan kesepuluhnya pada data pertemanan Facebook nyata dengan validasi silang Monte Carlo sehingga setiap angka dilengkapi simpangan baku dan selang kepercayaan, bukan hasil satu kali percobaan. Ketiga, menguji kestabilan peringkat melalui analisis sensitivitas terhadap rasio data uji serta uji tanda antarpasangan metrik. Keempat, menunjukkan secara analitis sekaligus empiris bahwa dua metrik yang tampak berbeda, yaitu Jaccard dan Sørensen, sesungguhnya ekuivalen-peringkat.

Sisa makalah disusun sebagai berikut. Bagian II meninjau pekerjaan terdahulu yang serupa dan menegaskan posisi tulisan

ini secara jujur. Bagian III memaparkan landasan teori graf dan himpunan yang diperlukan. Bagian IV menurunkan kesepuluh metrik beserta klasifikasi, contoh perhitungan tangan, dan hubungan analitis antarmetrik. Bagian V menjelaskan dataset, algoritma, kompleksitas, dan protokol pengujian Monte Carlo. Bagian VI menyajikan dan membahas hasil eksperimen, lalu Bagian VII menutup dengan kesimpulan dan arah pengembangan.

II. PEKERJAAN TERKAIT DAN POSISI MAKALAH

Membandingkan beberapa metrik kemiripan lokal untuk prediksi tautan adalah praktik yang lazim dalam riset jaringan kompleks. Liben-Nowell dan Kleinberg [3] memformalkan persoalan prediksi tautan dan menguji sejumlah penyeor pada jaringan kolaborasi ilmiah. Zhou, Lü, dan Zhang [8] membandingkan sembilan metrik lokal pada enam jaringan nyata dan memperkenalkan indeks Resource Allocation, sedangkan Lü dan Zhou [7] merangkum perkembangan bidang ini secara menyeluruh. Dengan demikian, makalah ini tidak mengklaim sebagai perbandingan yang pertama. Sebaliknya, ia adalah studi reproduksi sekaligus penerapan yang menempatkan gagasan tersebut pada kerangka Matematika Diskrit dan menambahkan kedisiplinan statistik pada pengujiannya.

Pada lingkup makalah Matematika Diskrit di lingkungan kampus, beberapa tulisan terdahulu telah menyinggung tema yang berdekatan. Harjono [14] menjelaskan cara kerja fitur friend suggestion dengan menekankan konsep teman bersama secara naratif. Yahya [15] membahas penentuan rekomendasi teman melalui penelusuran graf pertemanan. Fairuz [16] memodelkan hubungan pertemanan Facebook sebagai graf tak-berarah dan menguraikan sifat strukturalnya. Ketiga tulisan tersebut bersifat deskriptif: fokusnya menjelaskan bahwa graf dapat dipakai untuk merekomendasikan teman, tanpa merumuskan beberapa metrik berbeda, membandingkannya, atau menguji akurasinya pada data nyata.

Berdasarkan kedua kelompok rujukan tersebut, posisi makalah ini diletakkan pada dua hal yang dapat dipertanggungjawabkan. Pertama, seluruh metrik diturunkan secara mandiri dari operasi himpunan dan teori graf dasar, lalu diverifikasi pada contoh yang dapat dihitung tangan, sehingga selaras dengan tujuan pembelajaran Matematika Diskrit. Kedua, pengujian dilengkapi kerangka statistik berupa validasi silang Monte Carlo, selang kepercayaan, uji tanda, dan analisis sensitivitas, yang memberi dasar lebih kuat bagi setiap kesimpulan dibanding sekadar satu kali percobaan. Singkatnya, makalah ini memadukan ketelitian penurunan ala pembelajaran dengan ketelitian pengujian ala riset.

III. LANDASAN TEORI

A. Graf dan Himpunan Tetangga

Graf $G = (V, E)$ adalah struktur yang terdiri atas himpunan tak-kosong simpul V dan himpunan sisi E , dengan setiap sisi menghubungkan sepasang simpul [1], [2]. Pada makalah ini digunakan graf tak-berarah sederhana, yaitu graf tanpa sisi ganda dan tanpa gelang. Pemodelan ini sesuai dengan pertemanan yang bersifat timbal balik: jika u berteman dengan

v , maka v juga berteman dengan u , sehingga sisi $\{u, v\}$ tidak memiliki arah.

Dua simpul u dan v dikatakan bertetangga jika terdapat sisi $\{u, v\} \in E$. Himpunan semua simpul yang bertetangga dengan v disebut himpunan tetangga, dilambangkan $N(v)$. Banyaknya tetangga sebuah simpul disebut derajat, ditulis $k(v) = |N(v)|$. Dalam konteks jejaring sosial, $N(v)$ adalah himpunan teman pengguna v , dan $k(v)$ adalah jumlah temannya. Karena jejaring sosial umumnya jarang, yakni berisi jauh lebih sedikit sisi daripada batas maksimum $|V|(|V|-1)/2$, graf direpresentasikan sebagai senarai ketetanggaan, yaitu pemetaan dari setiap simpul ke himpunan tetangganya. Representasi ini cocok dengan kebutuhan makalah karena operasi irisan dan gabungan yang menjadi inti seluruh metrik dapat langsung dikenakan pada himpunan tetangga.

B. Operasi Himpunan

Himpunan adalah kumpulan objek yang terdefinisi dengan jelas [1]. Dua operasi yang menjadi inti makalah ini adalah irisan dan gabungan. Irisan $A \cap B$ adalah himpunan anggota yang terdapat pada A sekaligus pada B , sedangkan gabungan $A \cup B$ adalah himpunan anggota yang terdapat pada A atau B atau keduanya. Kardinalitas $|A|$ menyatakan banyaknya anggota A . Bila kedua operasi ini dikenakan pada himpunan tetangga dua simpul, $N(u) \cap N(v)$ tepat menyatakan himpunan teman bersama u dan v , sedangkan $N(u) \cup N(v)$ menyatakan keseluruhan teman keduanya tanpa pengulangan. Seluruh metrik yang dibahas berikutnya dibangun dari kedua besaran sederhana ini.

C. Notasi

Agar rumus pada bagian berikutnya ringkas, Tabel I merangkum notasi yang dipakai sepanjang makalah.

Tabel I. Notasi yang digunakan.

Simbol	Makna
$G = (V, E)$	graf, himpunan simpul dan sisi
$N(v)$	himpunan tetangga (teman) simpul v
$k(v) = N(v) $	derajat simpul v
c	banyak teman bersama, $ N(u) \cap N(v) $
$N(u) \cup N(v)$	gabungan teman u dan v
w	teman bersama, $w \in N(u) \cap N(v)$
$S(u, v)$	skor kemiripan pasangan (u, v)

D. Prediksi Tautan, Lintasan, dan Ruang Kandidat

Prediksi tautan adalah persoalan memperkirakan sisi yang belum ada namun berpeluang muncul pada sebuah graf [3]. Untuk setiap pasangan simpul (u, v) yang belum bertetangga didefinisikan fungsi skor $S(u, v)$ yang menyatakan seberapa besar kemungkinan keduanya akan terhubung; pasangan berskor tertinggi dijadikan rekomendasi. Dengan demikian, fitur “Orang yang Mungkin Anda Kenal” setara dengan menghitung $S(u, v)$ untuk pasangan kandidat lalu mengurutkannya menurun.

Sebuah lintasan pada graf adalah barisan simpul berurutan yang setiap dua simpul berurutannya dihubungkan sebuah sisi, dan jarak antara dua simpul adalah panjang lintasan terpendek yang menghubungkannya [1]. Pasangan (u, v) memiliki teman bersama, yakni $N(u) \cap N(v) \neq \emptyset$, jika dan hanya jika ada simpul w yang bertetangga dengan u sekaligus dengan v , yang berarti u dan v terhubung melalui lintasan $u-w-v$ sepanjang dua sisi. Jadi himpunan pasangan tak-bertetangga yang berjarak tepat dua sama persis dengan himpunan pasangan yang memiliki teman bersama. Inilah ruang kandidat yang ditinjau, karena di luar ruang itu seluruh metrik berbasis irisan pasti bernilai nol dan tidak memberikan informasi.

IV. PERUMUSAN DAN KLASIFIKASI METRIK

Seluruh metrik berikut hanya bergantung pada himpunan tetangga $N(u)$, $N(v)$ dan derajat $k(u)$, $k(v)$, sehingga selaras dengan landasan graf dan himpunan di atas. Untuk setiap pasangan (u, v) tuliskan $c = |N(u) \cap N(v)|$ sebagai banyak teman bersama. Kesepuluh metrik dikelompokkan menjadi tiga keluarga menurut cara mereka mengolah besaran ini.

A. Keluarga Berbasis Cacah

Keluarga ini memakai besaran mentah tanpa normalisasi. Common Neighbors (CN) menghitung langsung banyak teman bersama, sedangkan Preferential Attachment (PA) sama sekali tidak memakai irisan melainkan hasil kali derajat kedua simpul, bersandar pada gagasan bahwa simpul berderajat besar cenderung memperoleh tautan baru [3], [11]:

$$S_{CN}(u, v) = c = |N(u) \cap N(v)|,$$

$$S_{PA}(u, v) = k(u) \cdot k(v).$$

PA disertakan sebagai basis pembanding justru karena mengabaikan teman bersama; kinerjanya menjadi tolok ukur seberapa besar nilai tambah informasi irisan.

B. Keluarga Ternormalisasi

Keluarga ini membagi c dengan suatu ukuran kedua himpunan tetangga sehingga skornya berada pada $[0, 1]$ dan tak berdimensi. Koefisien Jaccard membagi dengan kardinalitas gabungan; indeks Salton (kosinus) membagi dengan rata-rata geometrik derajat [12]; indeks Sørensen membagi dengan rata-rata aritmetik derajat [13]; Hub Promoted Index (HPI) dan Hub Depressed Index (HDI) menormalkan terhadap derajat terkecil dan terbesar [10]; sedangkan indeks Leicht-Holme-Newman (LHN) membagi dengan hasil kali derajat [9]:

$$S_{Jaccard} = c / |N(u) \cup N(v)|,$$

$$S_{Salton} = c / \sqrt{k(u) \cdot k(v)},$$

$$S_{Sørensen} = 2c / (k(u) + k(v)),$$

$$S_{HPI} = c / \min(k(u), k(v)),$$

$$S_{HDI} = c / \max(k(u), k(v)),$$

$$S_{LHN} = c / (k(u) \cdot k(v)).$$

C. Keluarga Berbobot Derajat

Alih-alih memperlakukan setiap teman bersama setara, keluarga ini membobotnya menurut derajatnya: teman bersama yang sendirinya berteman sedikit dianggap lebih informatif daripada yang sangat populer. Resource Allocation

(RA) [8] memberi bobot $1/k(w)$, sedangkan Adamic-Adar (AA) [4] memberi bobot $1/\log k(w)$:

$$S_{RA} = \sum 1/k(w), \quad S_{AA} = \sum 1/\log k(w), \quad w \in N(u) \cap N(v).$$

Keduanya meredam pengaruh teman bersama berderajat besar; RA meredam lebih tajam karena $1/k(w)$ menurun lebih cepat daripada $1/\log k(w)$. Logaritma yang dipakai adalah logaritma natural; pilihan basis hanya menskala skor AA dengan konstanta sehingga tidak mengubah peringkat maupun AUC. Karena setiap teman bersama w pada graf latih pasti terhubung ke u dan v sekaligus, derajatnya selalu $k(w) \geq 2$; akibatnya $1/k(w)$ dan $1/\log k(w)$ senantiasa terdefinisi tanpa risiko pembagian oleh nol maupun $\log 1 = 0$.

D. Ringkasan Klasifikasi

Tabel II merangkum keluarga dan sifat pokok kesepuluh metrik. Kolom rentang menunjukkan batas nilai, dan kolom pembobotan menandai apakah tiap teman bersama dibobot menurut derajatnya.

Tabel II. Klasifikasi dan sifat kesepuluh metrik.

Metrik	Keluarga	Rentang	Bobot
CN	cacah	$[0, \infty)$	tidak
PA	cacah	$[0, \infty)$	tidak
Jaccard	ternormalisasi	$[0, 1]$	tidak
Salton	ternormalisasi	$[0, 1]$	tidak
Sørensen	ternormalisasi	$[0, 1]$	tidak
HPI	ternormalisasi	$[0, 1]$	tidak
HDI	ternormalisasi	$[0, 1]$	tidak
LHN	ternormalisasi	$[0, 1]$	tidak
RA	berbobot	$[0, \infty)$	ya
AA	berbobot	$[0, \infty)$	ya

Tanda $[0, \infty)$ pada kolom rentang menyatakan bahwa metrik tidak dinormalkan ke $[0, 1]$, bukan berarti nilainya benar-benar tak hingga; secara praktis nilainya tetap terbatas, misalnya $CN \leq \min(k(u), k(v))$ dan $PA \leq (|V| - 1)^2$.

E. Contoh Perhitungan

Pasangan (A, B) yang belum berteman;
C dan D adalah teman bersama (irisan tetangga)

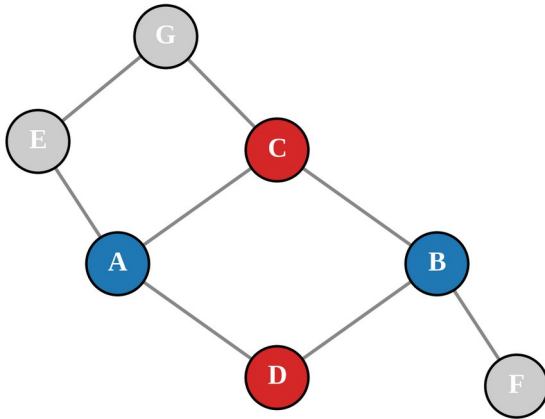


Fig. 1. Gambar 1. Graf contoh. A dan B belum berteman; C dan D adalah teman bersama mereka.

Tinjau graf pada Gambar 1 dengan pasangan dinilai (A, B) yang belum bertetangga. Diperoleh $N(A) = \{C, D, E\}$ dan $N(B) = \{C, D, F\}$, sehingga $c = |\{C, D\}| = 2$, gabungannya $\{C, D, E, F\}$ berukuran 4, $k(A) = k(B) = 3$, serta derajat teman bersama $k(C) = 3$ dan $k(D) = 2$. Substitusi ke kesepuluh rumus memberi:

$$S_{CN} = 2; S_{PA} = 9; S_{Jaccard} = 2/4 = 0,500;$$

$$S_{Salton} = S_{Sørensen} = S_{HPI} = S_{HDI} = 2/3 \approx 0,667;$$

$$S_{LHN} = 2/9 \approx 0,222; S_{RA} = 1/3 + 1/2 = 0,833;$$

$$S_{AA} = 1/\log 3 + 1/\log 2 \approx 2,353.$$

Karena pada contoh ini $k(A) = k(B)$, keempat metrik Salton, Sørensen, HPI, dan HDI kebetulan bernilai sama; perbedaannya baru tampak ketika derajat kedua simpul berbeda. Seluruh angka ini identik dengan keluaran program, sehingga implementasi terverifikasi pada kasus yang dapat dihitung tangan.

F. Hubungan Analitis Antarmetrik

Walaupun bersumber dari besaran yang sama, kesepuluh metrik memperlakukannya berbeda, dan beberapa hubungan dapat ditunjukkan secara analitis. Yang paling mencolok, Jaccard dan Sørensen ekuivalen-peringkat. Tuliskan $g = |N(u) \cup N(v)| = k(u) + k(v) - c$ sehingga $S_{Jaccard} = c/g$. Karena $k(u) + k(v) = g + c$, indeks Sørensen dapat dinyatakan sebagai fungsi murni dari $S_{Jaccard}$:

$$S_{Sørensen} = 2c / (g + c) = 2 \cdot S_{Jaccard} / (1 + S_{Jaccard}).$$

Fungsi $f(x) = 2x/(1 + x)$ naik tegas untuk $x \geq 0$ karena turunannya $f'(x) = 2/(1 + x)^2$ selalu positif. Akibatnya, urutan pasangan menurut Sørensen pasti sama dengan urutannya menurut Jaccard, sehingga keduanya menghasilkan AUC dan Precision@k yang identik berapa pun datanya. Sifat ini akan terlihat jelas pada hasil eksperimen dan berguna sebagai pemeriksaan silang kebenaran implementasi.

Hubungan lain juga dapat dicatat. LHN sama dengan $c/(k(u) \cdot k(v))$, yaitu CN dibagi PA, sehingga sangat menghukum pasangan berderajat besar. RA dan AA tereduksi menjadi kelipatan CN bila semua teman bersama berderajat sama, dan baru berbeda ketika derajat teman bersama beragam. Dengan demikian perbedaan kinerja antarmetrik sepenuhnya bersumber dari cara masing-masing memperlakukan derajat, bukan dari informasi irisan yang berbeda.

V. IMPLEMENTASI DAN RANCANGAN EKSPERIMEN

A. Dataset

Eksperimen menggunakan dataset ego-Facebook dari Stanford Network Analysis Project (SNAP) [5], [6], yang berisi data pertemanan Facebook anonim nyata. Graf ini memiliki 4.039 simpul (pengguna) dan 88.234 sisi (pertemanan), dengan derajat rata-rata 43,69 dan derajat maksimum 1.045. Distribusi derajatnya (Gambar 2) memperlihatkan pola berekor panjang yang khas jejaring sosial nyata: sebagian besar pengguna memiliki sedikit teman, sementara sejumlah kecil pengguna sangat terhubung.

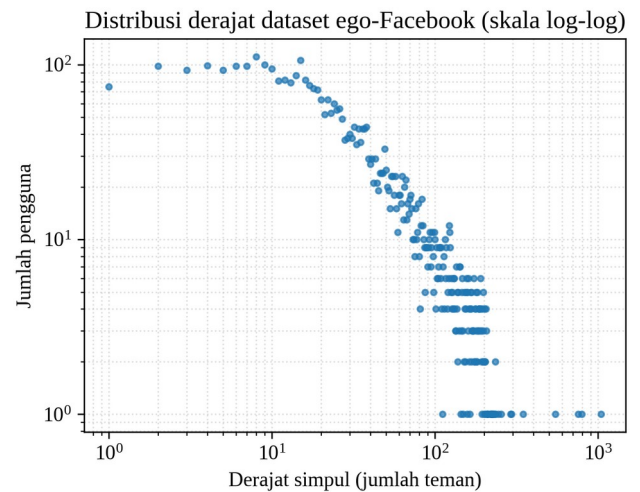


Fig. 2. Gambar 2. Distribusi derajat dataset ego-Facebook pada skala log-log.

B. Validasi Silang Monte Carlo

Mengukur akurasi dari satu kali pemisahan data uji rentan terhadap kebetulan: sisi mana yang terpilih untuk disembunyikan dapat menggeser angka, dan satu percobaan tidak memberi cara menilai apakah selisih antarmetrik bersifat nyata. Untuk itu digunakan validasi silang Monte Carlo, yaitu mengulang seluruh prosedur sebanyak 20 kali, masing-masing dengan pemisahan acak 80% sisi sebagai graf latih dan 20% sebagai data uji memakai benih acak berbeda. Hasil tiap metrik dilaporkan sebagai rerata dan simpangan baku 20 ulangan, dilengkapi selang kepercayaan 95% berdasarkan hampiran normal; dengan 20 ulangan, selang berbasis distribusi-t hanya sedikit lebih lebar dan tidak mengubah satu pun kesimpulan. Simpangan baku yang dilaporkan adalah simpangan baku populasi atas 20 ulangan; memakai penduga sampel berpembagi $n-1$ hanya melebarkan nilainya sekitar 2,6% dan tidak mengubah kesimpulan. Dengan demikian, pernyataan bahwa satu metrik lebih baik daripada yang lain dapat diuji, bukan sekadar diklaim.

C. Ruang Kandidat

Seluruh skor dihitung hanya dari graf latih, lalu metrik diuji pada kemampuannya menemukan kembali sisi yang disembunyikan. Ruang kandidat dibatasi pada pasangan teman dari teman, yaitu pasangan tak-bertetangga yang memiliki minimal satu teman bersama pada graf latih. Pembatasan ini realistis karena fitur rekomendasi memang hanya menelusuri tetangga sejauh dua langkah, sekaligus efisien karena pasangan tanpa teman bersama pasti berskor nol. Dari proses ini diperoleh sekitar 994.774 pasangan kandidat pada satu ulangan representatif, di antaranya sekitar 17.495 pasangan positif berupa pertemanan yang disembunyikan dan sisanya negatif, yaitu pasangan yang bukan sisi pada graf penuh sehingga bukan sisi latih maupun sisi uji; jumlah kandidat berdekatan antar-ulangan karena pemisahan acak hanya menggeser sedikit susunan graf latih. Sebagai konsekuensi pembatasan ini, sebagian kecil sisi uji yang kedua simpulnya tidak lagi memiliki teman bersama pada graf latih (sekitar 0,9% pada rasio uji 20%) berada di luar ruang kandidat sehingga tidak ikut dinilai; dengan kata lain, kinerja diukur atas sisi yang memang terjangkau dalam dua langkah, konsisten dengan cakupan fitur rekomendasi yang dimodelkan.

D. Algoritma

Keseluruhan proses diringkas pada pseudokode berikut. Perulangan terluar adalah ulangan Monte Carlo; di dalamnya graf latih diubah menjadi pemetaan tetangga, ruang kandidat dibangkitkan, lalu setiap metrik menyekor dan dievaluasi.

Algoritma 1. Alur eksperimen prediksi tautan Monte Carlo.

MASUKAN : graf $G = (V, E)$, banyak ulangan R
KELUARAN: rerata dan simpangan baku AUC serta Precision@k tiap metrik

```
1  untuk  $r = 1$  sampai  $R$ :
2      pisah acak  $E \rightarrow E_{\text{latih}} (80\%), E_{\text{uji}} (20\%)$ 
3      bentuk  $N(v)$  untuk tiap  $v$  dari  $E_{\text{latih}}$ 
4       $K \leftarrow$  pasangan 'teman dari teman' bukan
        sisi latih // ruang kandidat
5      untuk setiap metrik  $S$ :
6          untuk setiap  $\{u, v\}$  dalam  $K$ :
7              skor  $\leftarrow S(u, v)$  memakai  $N(\cdot)$ 
8              label  $\leftarrow 1$  jika  $\{u, v\}$  di  $E_{\text{uji}}$ 
9              catat AUC dan Precision@k metrik  $S$ 
10     simpan hasil ulangan  $r$ 
11 laporkan rerata  $\pm$  simpangan baku antar-ulangan
```

E. Kompleksitas

Pembentukan himpunan tetangga berlangsung dalam waktu $O(|E|)$. Pembangkitan kandidat menyusuri tiap simpul w dan seluruh pasangan tetangganya, sehingga biayanya sebanding dengan $\sum k(w)^2$, yang dalam praktik tetap terkendali karena distribusi derajat berekor panjang dan hanya sedikit simpul berderajat besar. Penyekoran satu pasangan berbiaya $O(\min(k(u), k(v)))$ karena cukup menelusuri himpunan tetangga yang lebih kecil. Dalam implementasi, kesepuluh metrik dihitung sekaligus dalam satu kali penelusuran irisan, sehingga menambah metrik nyaris tidak menambah biaya.

F. Metrik Evaluasi dan Uji Tanda

Dua ukuran baku digunakan. Pertama, AUC (Area Under the ROC Curve), yang dapat ditafsirkan sebagai peluang sebuah pasangan positif acak memperoleh skor lebih tinggi

daripada pasangan negatif acak; nilainya 0,5 berarti setara tebakan acak dan 1,0 berarti sempurna. AUC dihitung melalui metode peringkat yang setara dengan statistik U Mann-Whitney. Kedua, Precision@k, yaitu proporsi pasangan positif di antara k rekomendasi berskor tertinggi, yang mencerminkan pengalaman pengguna yang hanya melihat sedikit saran teratas.

Untuk menilai keajekan urutan dua metrik digunakan uji tanda sederhana: dihitung pada berapa dari 20 ulangan suatu metrik mengungguli metrik lain. Bila satu metrik menang pada seluruh 20 ulangan, peluang kejadian itu di bawah hipotesis nol bahwa kedua metrik setara hanyalah $(1/2)^{20} \approx 9,5 \times 10^{-7}$ (uji tanda satu sisi), sehingga keunggulannya dapat dianggap konsisten, bukan kebetulan. Perlu dicatat bahwa kedua puluh ulangan berbagi graf dasar yang sama sehingga pemisahannya tidak sepenuhnya saling bebas; nilai p tersebut karenanya bersifat hampiran yang cenderung optimistis. Namun karena selisih antarmetrik jauh lebih besar daripada simpangan bakunya, kesimpulan keajekan peringkat tetap berlaku.

G. Lingkungan

Seluruh program ditulis dalam bahasa Python tanpa pustaka komputasi graf khusus; graf direpresentasikan sebagai pemetaan dari simpul ke himpunan tetangga sehingga operasi irisan dan gabungan langsung memakai tipe data himpunan bawaan. Setiap ulangan Monte Carlo selesai dalam belasan detik pada komputer biasa, menegaskan bahwa pendekatan ini ringan dan tidak memerlukan perangkat keras khusus.

VI. HASIL DAN PEMBAHASAN

Tabel III memuat AUC rerata dan simpangan baku kesepuluh metrik dari 20 ulangan Monte Carlo, terurut dari yang terbaik. Resource Allocation menempati posisi teratas (0,972), disusul Adamic-Adar (0,962) dan Common Neighbors (0,952). Simpangan baku seluruh metrik sangat kecil, di bawah 0,004, sehingga selang kepercayaan antarmetrik yang berbeda peringkat tidak saling tumpang tindih, dengan satu pengecualian wajar, yaitu Jaccard dan Sørensen yang ekuivalen-peringkat sehingga AUC-nya memang identik. Gambar 3 menampilkan perbandingan ini lengkap dengan batang galat yang memperlihatkan betapa rapatnya sebaran antar-ulangan.

Tabel III. AUC rerata dan simpangan baku (20 ulangan).

Metrik	AUC	Simpangan baku
RA	0,9718	0,0007
AA	0,9619	0,0008
CN	0,9520	0,0010
Salton	0,9450	0,0012
Jaccard	0,9360	0,0012
Sørensen	0,9360	0,0012
HDI	0,9293	0,0013
HPI	0,9070	0,0028
PA	0,7835	0,0021
LHN	0,6976	0,0035

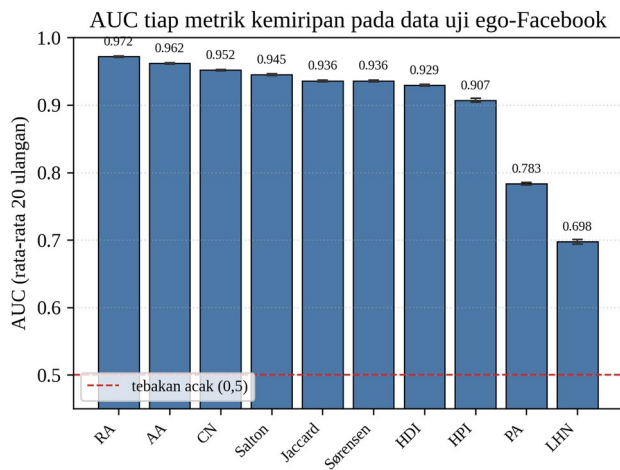


Fig. 3. Gambar 3. AUC tiap metrik dengan batang galat satu simpangan baku; garis putus-putus menandai tebakan acak (0,5).

Uji tanda antarpasangan metrik bersebelahan memperkuat kesimpulan ini. Dari sembilan pasangan metrik bersebelahan, pada delapan pasangan metrik berperingkat lebih tinggi menang di seluruh 20 ulangan tanpa kecuali. Satu-satunya pengecualian adalah pasangan Jaccard dan Sørensen, yang tidak pernah saling mengungguli pada satu ulangan pun (0 dari 20). Hasil ini persis seperti yang diramalkan pada Bagian IV-F: karena keduanya ekuivalen-peringkat, AUC-nya identik hingga digit yang dilaporkan, yaitu 0,9360, dan tidak ada ulangan yang dapat memisahkannya. Kesesuaian antara prediksi analitis dan pengamatan empiris ini menjadi pemeriksaan silang yang meyakinkan atas kebenaran implementasi.

Tabel IV dan Gambar 4 menyajikan gambaran yang lebih tajam. Tiga metrik ber-AUC tertinggi, yaitu RA, AA, dan CN, juga mempertahankan presisi sangat tinggi di sepanjang daftar, di atas 0,90 hingga $k = 1.000$. Sebaliknya, keluarga ternormalisasi yang AUC-nya tergolong baik justru memiliki presisi amat rendah pada sedikit rekomendasi teratas: Jaccard, Salton, Sørensen, dan HDI hanya mencapai sekitar 0,035 pada $k = 10$, yang berarti dari sepuluh saran teratas mereka hampir tidak ada yang benar.

Tabel IV. Precision@k tiap metrik.

Metrik	P@10	P@100	P@1000
RA	1,000	0,977	0,950
AA	1,000	0,981	0,919
CN	0,995	0,980	0,915
Salton	0,035	0,019	0,473
Jaccard	0,035	0,019	0,477
Sørensen	0,035	0,019	0,477
HDI	0,035	0,019	0,506
HPI	0,015	0,005	0,008
PA	0,255	0,347	0,289
LHN	0,000	0,003	0,004

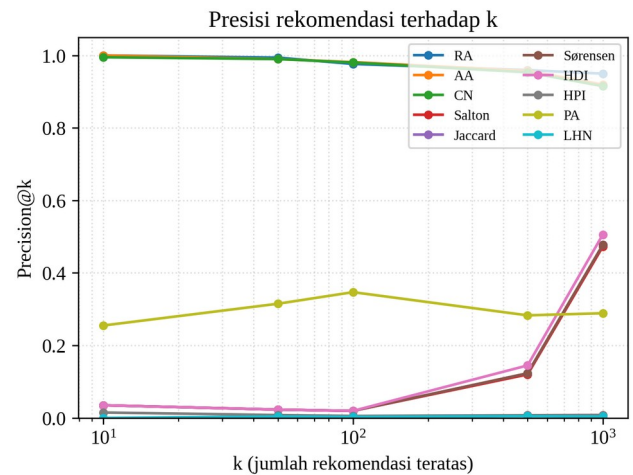


Fig. 4. Gambar 4. Precision@k tiap metrik terhadap panjang daftar rekomendasi k (sumbu k berskala logaritmik).

A. Mengapa Metrik Ternormalisasi Gagal di Puncak

Paradoks AUC tinggi tetapi presisi puncak rendah berakar pada perbedaan sifat kedua ukuran. AUC bersifat global: ia hanya menanyakan apakah pasangan positif cenderung berperingkat di atas pasangan negatif secara menyeluruh. Precision@k bersifat lokal pada puncak daftar. Metrik ternormalisasi memberi skor maksimum kepada pasangan berderajat sangat kecil yang seluruh tetangganya kebetulan sama; misalnya dua pengguna yang masing-masing hanya memiliki satu teman dan teman itu sama akan memperoleh Jaccard = 1. Pasangan langka semacam itu membanjiri puncak daftar padahal jarang merupakan pertemanan sungguhan, sehingga presisi pada k kecil anjlok. Pada k besar, ketika kandidat terbukti banyak teman bersama akhirnya ikut terangkat, presisinya membaik kembali, terlihat dari kenaikan menuju $k = 1.000$. Metrik berbasis cacah dan berbobot tidak mengalami hal ini karena tetap menghargai banyaknya bukti, sehingga pasangan tipis berderajat satu tak pernah menggeser pasangan terbukti kuat dari puncak. Perlu ditambahkan satu catatan metodologis: pada metrik ternormalisasi puncak daftar dipenuhi banyak pasangan berskor seri sempurna bernilai 1, sehingga nilai Precision@k pada k kecil bergantung pada cara pemecahan seri dan sebaiknya dibaca sebagai besaran kasar yang sangat rendah, bukan ditafsirkan setajam digit ketiganya.

B. Keunggulan Pembobotan Kelangkaan

Di antara metrik berbobot, keunggulan tipis namun konsisten RA atas AA memperlihatkan manfaat pembobotan kelangkaan yang lebih tegas. Karena bobot $1/k(w)$ menekan teman bersama berderajat besar lebih kuat daripada $1/\log k(w)$, RA lebih jeli menyaring teman bersama yang sekadar populer. Menariknya, Zhou, Lü, dan Zhang [8] pada enam jaringan lain justru menemukan Common Neighbors unggul secara keseluruhan dengan RA sebagai pendatang yang menjanjikan; pada data ego-Facebook ini RA jelas terdepan. Perbedaan ini menegaskan bahwa peringkat persisnya bergantung pada karakteristik jaringan, sehingga analisis sensitivitas pada subbagian berikut menjadi penting.

C. Analisis Sensitivitas

Untuk menguji apakah peringkat bertahan di luar pilihan rasio uji 20%, eksperimen diulang pada rasio 10% dan 30% (Tabel V). Urutan kesepuluh metrik praktis tidak berubah: RA tetap teratas pada ketiga rasio, diikuti AA dan CN, sementara LHN selalu di dasar. Seluruh AUC sedikit menurun ketika data latih dikurangi pada rasio uji 30%, karena graf latih yang lebih jarang menyediakan lebih sedikit teman bersama, tetapi penurunannya seragam dan tidak mengubah kesimpulan apa pun.

Tabel V. AUC rerata pada tiga rasio data uji r (masing-masing 20 ulangan, sehingga kolom r = 0,20 setara dengan Tabel III).

Metrik	r=0,10	r=0,20	r=0,30
RA	0,976	0,972	0,966
AA	0,967	0,962	0,955
CN	0,960	0,952	0,942
Salton	0,956	0,945	0,928
Jaccard	0,947	0,936	0,921
Sørensen	0,947	0,936	0,921
HDI	0,940	0,929	0,914
HPI	0,923	0,907	0,883
PA	0,790	0,783	0,777
LHN	0,726	0,698	0,665

D. Contoh Rekomendasi

Sebagai ilustrasi penerapan, Tabel VI menampilkan lima rekomendasi teratas bagi pengguna bernomor 0 menurut Resource Allocation, dihitung dari graf latih. Kelima kandidat memiliki puluhan teman bersama dengan pengguna 0 meskipun belum terhubung, persis seperti yang diharapkan dari fitur rekomendasi yang baik.

Tabel VI. Lima rekomendasi teratas (RA) bagi pengguna 0.

Kandidat	Skor RA	Teman bersama
322	2,646	47
277	2,420	40
252	1,784	38
285	1,749	28
312	1,624	15

Penerapannya pada produk nyata sederhana: telusuri teman dari teman seorang pengguna untuk memperoleh himpunan kandidat, hitung skor RA tiap kandidat, lalu tampilkan beberapa yang tertinggi. Seluruh langkah hanya melibatkan operasi himpunan dan tidak memerlukan pelatihan model, sehingga rekomendasi dapat diperbarui cepat setiap kali graf pertemanan berubah.

E. Keterbatasan dan Ancaman terhadap Validitas

Beberapa keterbatasan perlu dicatat. Pertama, evaluasi memakai satu dataset dari satu platform; meski analisis sensitivitas menunjukkan peringkat yang stabil terhadap rasio uji, besaran angka dapat berbeda pada jejaring lain,

sebagaimana terlihat dari perbedaan dengan temuan [8]. Perlu pula ditekankan bahwa nilai AUC bersifat kondisional terhadap ruang kandidat: karena pasangan negatif dibatasi pada yang memiliki sedikitnya satu teman bersama, AUC absolut yang tinggi sebagian disebabkan ruang uji yang dipersempit, sehingga perbandingan angka lintas-studi harus memperhatikan definisi ruang kandidat masing-masing. Kedua, sisi yang disembunyikan dianggap satu-satunya pertemanan yang benar, padahal pasangan berlabel negatif bisa saja merupakan pertemanan nyata yang belum tercatat, sehingga presisi terukur cenderung lebih rendah daripada sebenarnya. Ketiga, seluruh metrik hanya memanfaatkan tetangga berjarak dua dan mengabaikan lintasan lebih panjang maupun atribut pengguna. Kesederhanaan ini justru menjadi kekuatan dari sudut pandang pembelajaran: seluruh metode dapat diturunkan, dihitung tangan, dan diuji hanya dengan operasi himpunan.

VII. KESIMPULAN

Makalah ini menunjukkan bahwa fitur rekomendasi teman pada jejaring sosial dapat dirumuskan secara utuh memakai teori graf dan operasi himpunan, lalu dievaluasi secara ketat. Sepuluh metrik kemiripan diturunkan dari irisan dan gabungan himpunan tetangga, dikelompokkan menjadi keluarga berbasis cacah, ternormalisasi, dan berbobot derajat, kemudian dibandingkan pada data pertemanan Facebook nyata dengan validasi silang Monte Carlo. Resource Allocation terbukti paling akurat ($AUC\ 0,972 \pm 0,001$) dan, bersama Adamic-Adar serta Common Neighbors, mempertahankan presisi tinggi pada daftar rekomendasi yang panjang. Keluarga ternormalisasi mencapai AUC baik tetapi gagal di puncak daftar karena mengunggulkan pasangan berderajat sangat kecil. Disimpulkan bahwa memberi bobot lebih besar kepada teman bersama yang langka, sebagaimana dilakukan RA dan AA, menghasilkan rekomendasi paling tepat untuk kasus ini.

Selain hasil empiris, makalah ini menegaskan sebuah hubungan analitis: koefisien Jaccard dan Sørensen ekuivalen-peringkat sehingga selalu menghasilkan AUC identik. Kesesuaian prediksi ini dengan eksperimen memperlihatkan nilai memadukan penalaran matematis dengan pengujian, bukan mengandalkan salah satunya saja.

Pengembangan lanjutan dapat menjajaki metrik berbasis lintasan yang lebih panjang, penggabungan beberapa metrik menjadi skor majemuk, misalnya menyaring kandidat dengan ambang teman bersama minimum sebelum memeringkat sisanya dengan RA, serta pengujian pada jejaring dari platform lain untuk menilai keberlakuan temuan secara lebih luas.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada dosen dan asisten mata kuliah IF1220 Matematika Diskrit atas penyampaian materi teori graf dan teori himpunan yang menjadi landasan makalah ini, serta kepada Stanford Network Analysis Project yang menyediakan dataset ego-Facebook secara terbuka sehingga eksperimen pada data nyata dapat dilakukan.

KETERSEDIAAN KODE DAN DATA

Seluruh kode sumber, dataset, dan keluaran numerik yang digunakan untuk menghasilkan hasil pada makalah ini tersedia secara terbuka di <https://github.com/aniceswan/link-prediction->

facebook sehingga seluruh eksperimen dapat ditelusuri dan direproduksi.

REFERENSI

- [1] K. H. Rosen, Discrete Mathematics and Its Applications, 8th ed. New York, NY, USA: McGraw-Hill, 2019.
- [2] R. Munir, Matematika Diskrit, 5th ed. Bandung, Indonesia: Penerbit Informatika, 2016.
- [3] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," Journal of the American Society for Information Science and Technology, vol. 58, no. 7, pp. 1019–1031, 2007.
- [4] L. A. Adamic and E. Adar, "Friends and neighbors on the Web," Social Networks, vol. 25, no. 3, pp. 211–230, 2003.
- [5] J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks," in Advances in Neural Information Processing Systems (NIPS), 2012, pp. 539–547.
- [6] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford Large Network Dataset Collection," <http://snap.stanford.edu/data>, diakses 18 Juni 2026.
- [7] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," Physica A: Statistical Mechanics and its Applications, vol. 390, no. 6, pp. 1150–1170, 2011.
- [8] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," The European Physical Journal B, vol. 71, no. 4, pp. 623–630, 2009.
- [9] E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex similarity in networks," Physical Review E, vol. 73, no. 2, art. 026120, 2006.
- [10] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," Science, vol. 297, no. 5586, pp. 1551–1555, 2002.
- [11] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," Science, vol. 286, no. 5439, pp. 509–512, 1999.
- [12] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval. New York, NY, USA: McGraw-Hill, 1983.
- [13] T. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content," Kongelige Danske Videnskabernes Selskab, Biologiske Skrifter, vol. 5, no. 4, pp. 1–34, 1948.
- [14] O. M. Harjono, "Aplikasi Graf pada Fitur Friend Suggestion di Media Sosial," Makalah IF2120 Matematika Diskrit, Institut Teknologi Bandung, 2014. [Daring]. Tersedia: <https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2014-2015/Makalah2014/Makalah-IF2120-2014-028.pdf>, diakses 18 Juni 2026.
- [15] Yahya, "Aplikasi Teori Graf untuk Menentukan Rekomendasi Teman pada Media Sosial," Makalah IF2120 Matematika Diskrit, Institut Teknologi Bandung, 2019. [Daring]. Tersedia: <https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2019-2020/Makalah2019/13518029.pdf>, diakses 18 Juni 2026.
- [16] M. A. Fairuz, "Aplikasi Undirected Graph untuk Hubungan Pertemanan pada Facebook," Makalah IF1220 Matematika Diskrit, Institut Teknologi Bandung, 2023. [Daring]. Tersedia: [https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2023-2024/Makalah2023/Makalah-Matdis-2023%20\(26\).pdf](https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2023-2024/Makalah2023/Makalah-Matdis-2023%20(26).pdf), diakses 18 Juni 2026.

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 18 Juni 2026

Ttd



Nabil Rabbani - 13525004