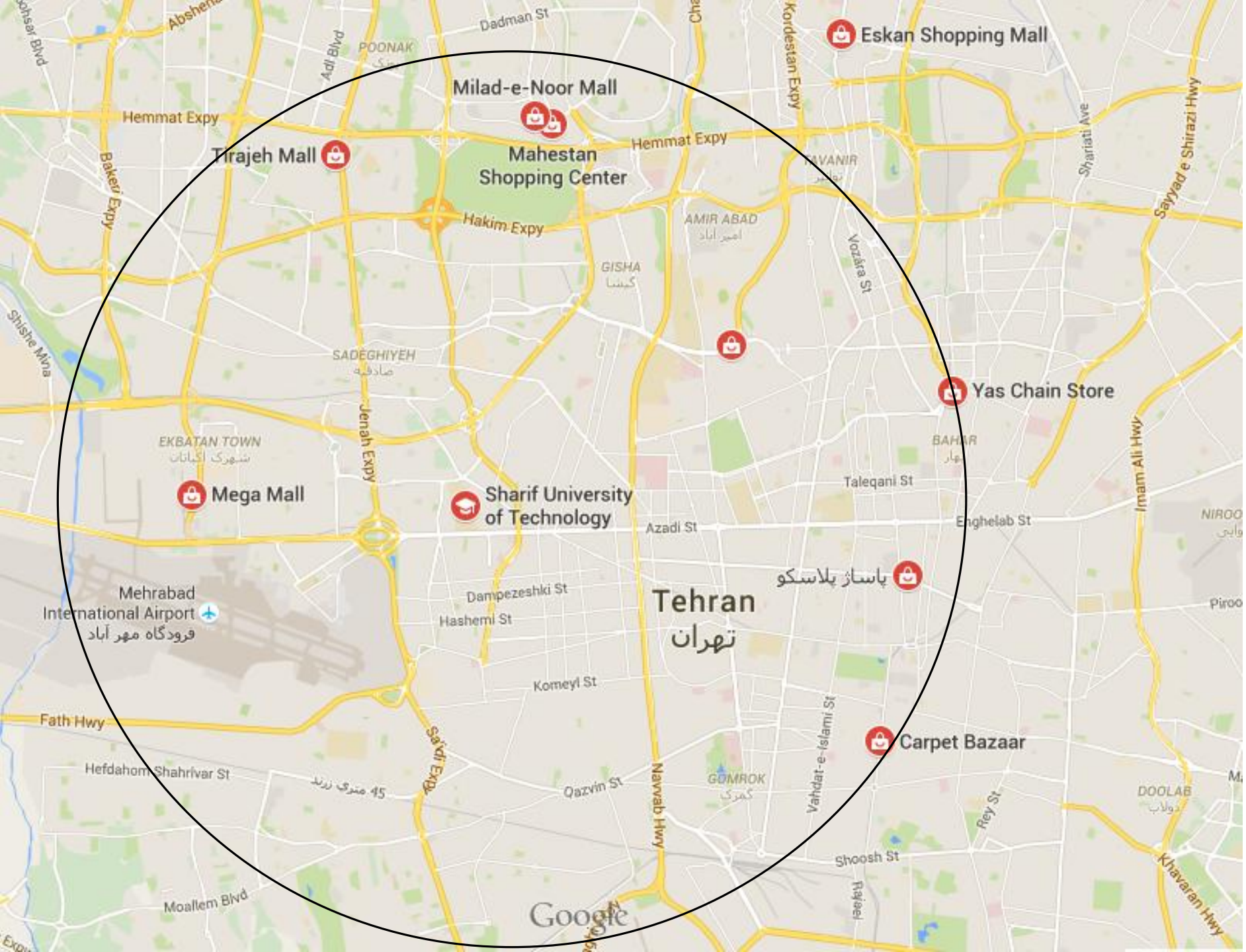
الگوریتم‌های تقریبی برای یافتن همسایه‌های نزدیک متنوع

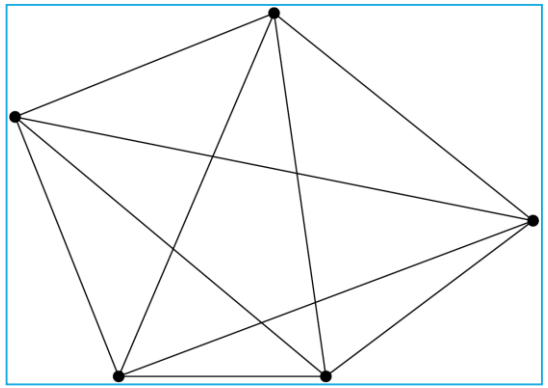
دانشجو: سپیده آقاملائی
استاد راهنما: دکتر حمید ضرابی زاده
استاد مشاور: دکتر محمد علی آبام
استاد مدعو: دکتر علیرضا زارعی

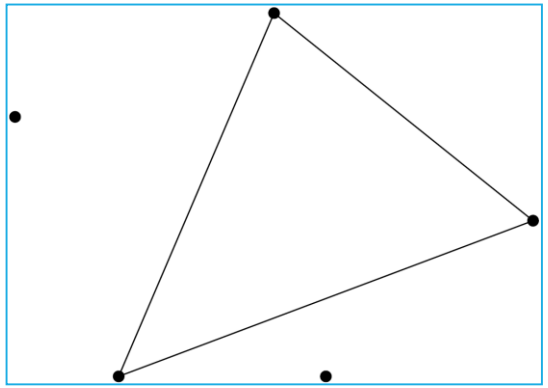


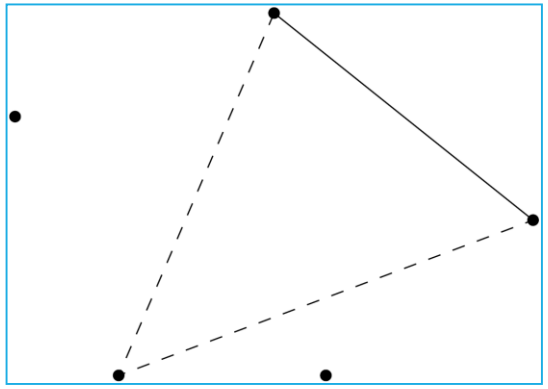
همسایه‌های نزدیک متنوع: مراکز خرید نزدیک دانشگاه

فهرست مطالب

- مسایل تنوع
- همسایه‌های نزدیک و درهم‌سازی حساس به محل
- همسایه‌های نزدیک متنوع
- مجموعه‌های هسته‌ی ترکیب‌شونده
- کران پایین مسایل تنوع در این مدل
- الگوریتم‌های گنزالز (حریصانه) و جستجوی محلی (برای تنوع)
- بهبود ضریب تقریب‌های قبلی
- تعریف جدید (مجموعه‌های هسته‌ی ترکیب‌شونده‌ی افرازی) و بهبود بیشتر ضریب تقریب‌ها
- نتیجه گیری و کارهای آینده







مسایل تنوع

ورودی

- مجموعه‌ی P شامل n نقطه
- عدد صحیح k
- ملاک تنوع: یک ویژگی π وابسته به توپولوژی گراف ساخته شده روی نقاط

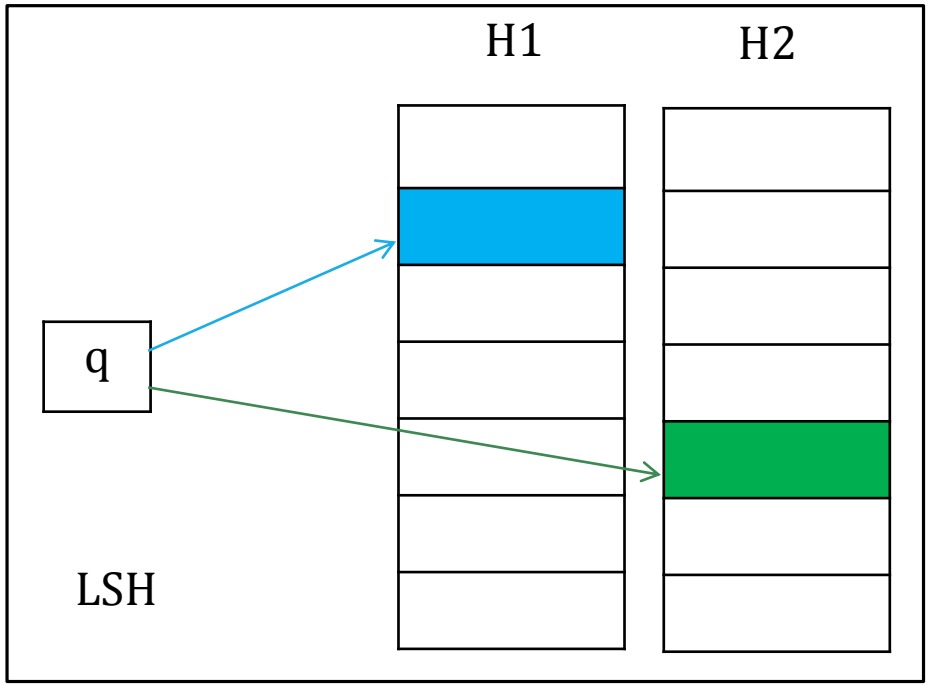
خروجی

- زیرمجموعه‌ای از k نقطه مانند Q که مقدار π بیشینه کند.

مثال‌هایی از π

- تنوع خوشه: جمع وزن یالهای مجموعه نقاط
- تنوع درخت: وزن درخت پوشای کمینه. اگر t درخت پوشای کمینه مجزا بخواهیم: t-درخت
- تنوع دور: وزن دور فروشنده‌ی دوره‌گرد. اگر t دور فروشنده دوره‌گرد مجزا بخواهیم: t-دور
- تنوع یال: وزن کوچکترین یال

همسایه‌های نزدیک



ورودی

- مجموعه نقاط P
- نقطه‌ی پرس‌وجوی q
- شعاع r

خروجی

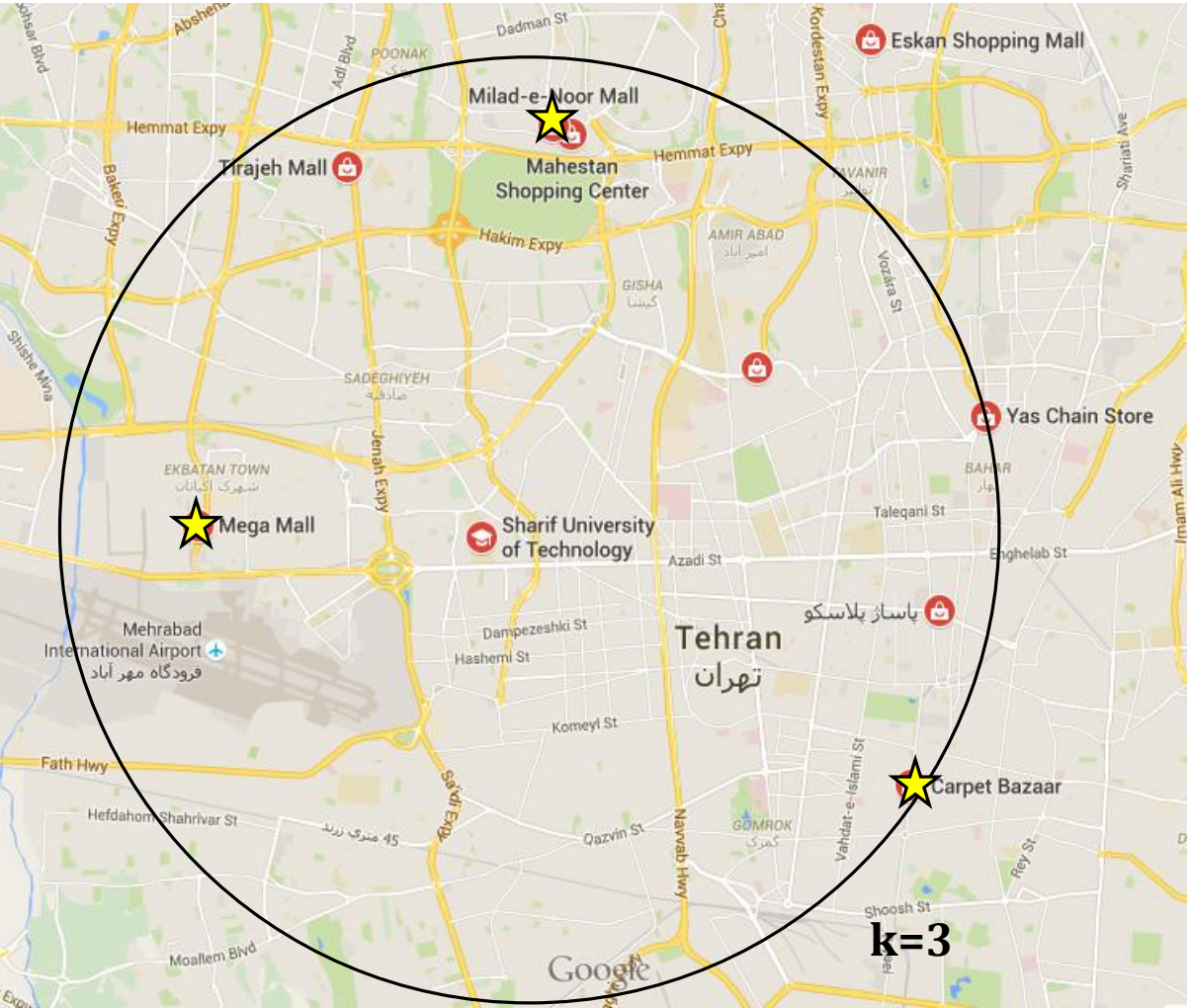
- تمام نقاطی که از q به فاصله‌ی حداکثر r هستند.

درهم‌سازی حساس به محل

- یک روش تصادفی تقریبی
- تابع درهم‌سازی که در آن احتمال نسبت دادن یک نقطه q به سطل حاوی نزدیک‌ترین همسایه‌ی تقریبی آن p بیشتر از احتمال نسبت دادن آن به سطل‌های دیگر باشد.

$$Pr[h(q)=h(p)] > Pr[h(q)\neq h(p)]$$

همسایه های نزدیک متنوع



ورودی

- مجموعه نقاط P
- نقطه پرس‌وجوی q
- شعاع r
- عدد صحیح k
- ملاک تنوع

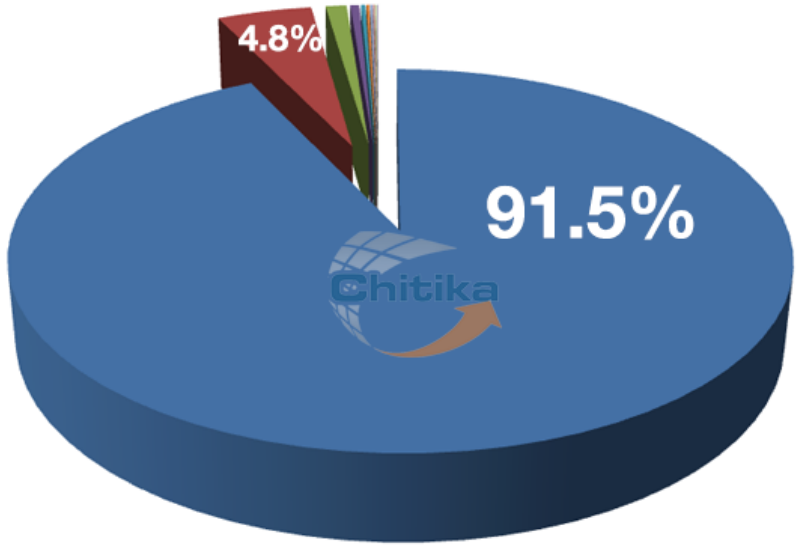
خروجی

- k نقطه که فاصله آنها تا q حداکثر r باشد و
- فاصله نسبی این نقاط نسبت به هم (تنوع نقاط) بیشینه شود.

کاربرد

- انتخاب نتایج موتور جستجو

Percentage of Google Traffic by Results Page



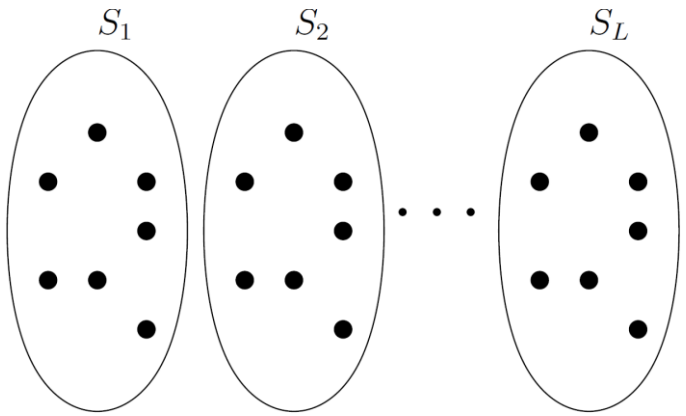
	Percentage of Google Traffic
Page 1	91.5%
Page 2	4.8%
Page 3	1.1%
Page 4	0.4%
Page 5	0.2%
Page 6	0.2%
Page 7	0.1%
Page 8	0.1%
Page 9	0.1%
Page 10	0.1%

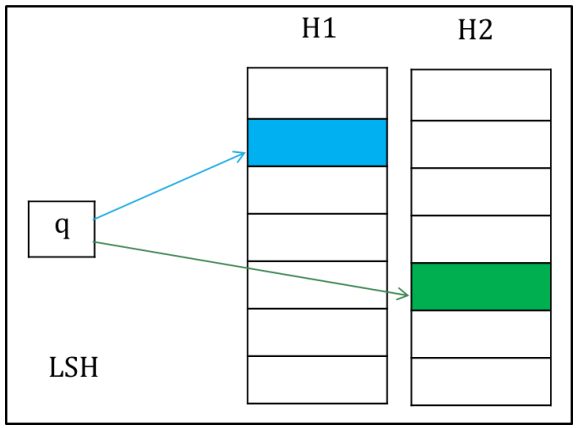
Data Driven by Chitika Insights June 2013
Data is representative of Google search traffic within the U.S. and Canada only

اهمیت موضوع

The best place to hide a dead body is page 2 of Google search results.

مجموعه‌های هسته‌ی ترکیب‌شونده





هدف: خلاصه‌سازی نقاط خروجی

- مجموعه‌ی هسته

چالش:

- داده‌های توزیع‌شده
- موازی‌سازی بدون تبادل داده

راهکار:

- یافتن راهی برای تبدیل خلاصه‌های هر مجموعه به یک خلاصه‌ی متمرکز

مثال:

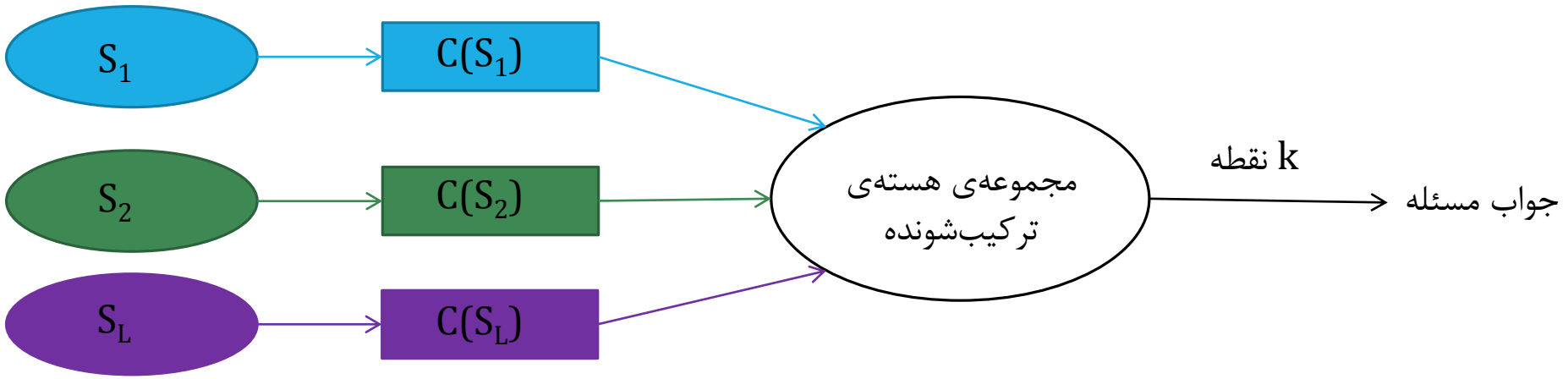
- هر سطل درهم‌سازی حساس به محل در هر تابع درهم‌سازی

مجموعه‌های هسته‌ی ترکیب‌شونده

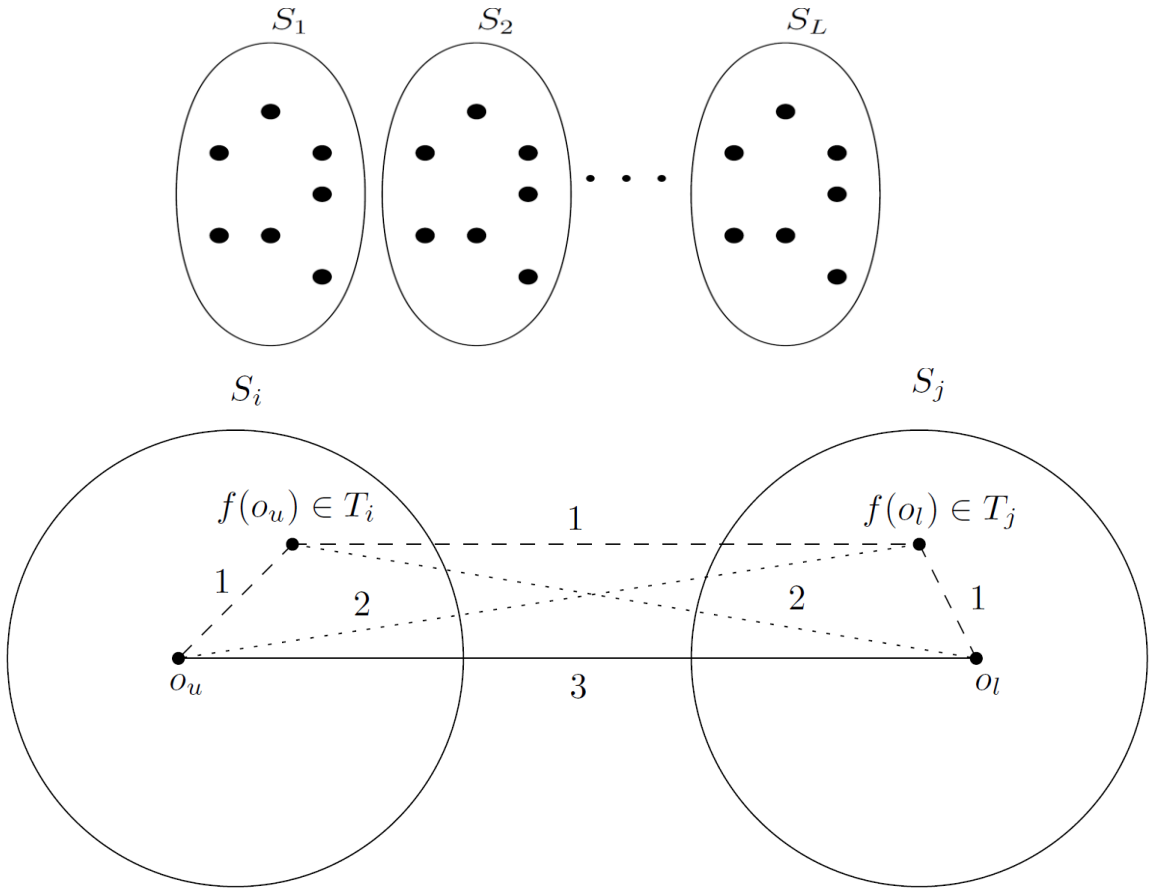
مجموعه‌های هسته‌ی ترکیب‌شونده

- مجموعه‌ی نقاط ورودی در مجموعه‌های $S_1,...,S_L$ توزیع شده‌اند و امکان تبادل اطلاعات ندارند و فقط از هر مجموعه $O(k)$ نقطه قابل اضافه کردن به مجموعه‌ی جواب است.
- می‌خواهیم k نقطه در مجموعه جواب بازگردانده شده باشد که تابع هدف (تنوع) را تقریب بزند.

$$div_k\ (S_1 \cup \ ... \cup S_L) \leq div_k\ (c(S_1) \cup \ ... \cup c(S_L))$$



کران پایین مسایل تنوع در این مدل



به ازای L=k مجموعه، فاصله‌ها را مطابق شکل قرار می‌دهیم.

کران پایین ضریب تقریب برای مسایل تنوع گفته شده ۳ به دست می‌آید.

نکات:

- درون هر مجموعه تقارن وجود داشته باشد. (نقاط بهینه و غیربهینه قابل تشخیص نباشند.)
- در گراف مجموعه‌ی هسته‌ی ترکیب‌شونده تقارن وجود داشته باشد.
- نامساوی مثلث برقرار بماند.

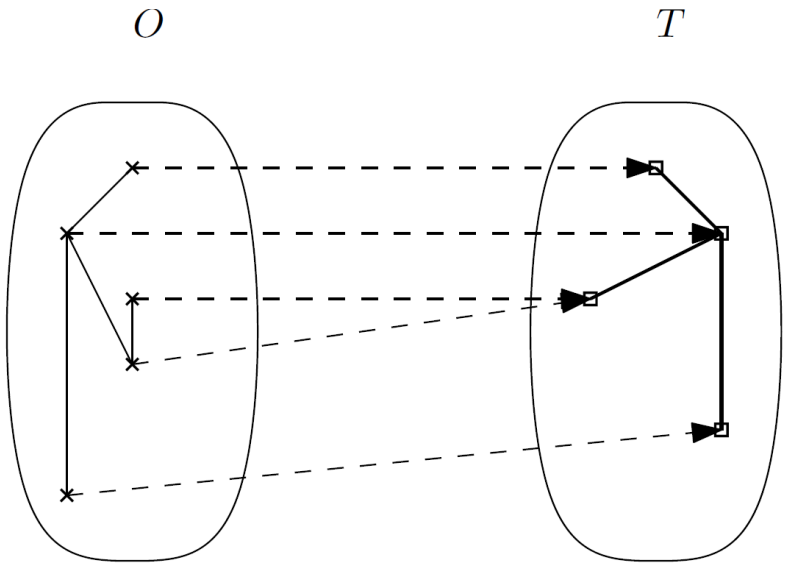
الگوریتم کلی و روند اثبات

(الگوریتم) ابتدا به ازای هر مجموعه‌ی S_i مجموعه‌ی هسته‌ی تنوع برای آن یعنی $\mathbf{c(S_i)}$ را محاسبه می‌کنیم.

(اثبات) سپس یک **تناظر** بین نقاط جواب بهینه و یک جواب روی مجموعه‌ی هسته (T) پیدا می‌کنیم.

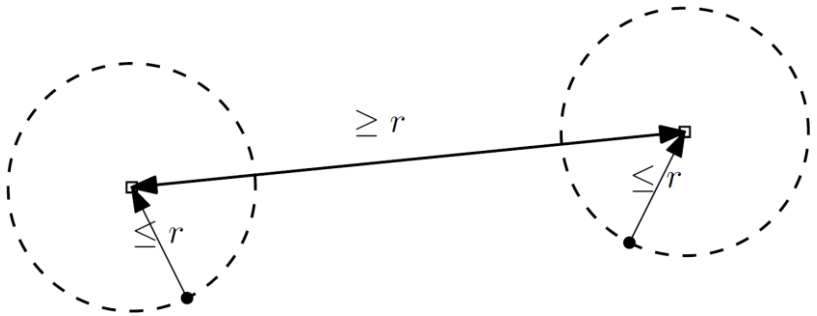
(الگوریتم) **اجتماع** مجموعه‌های هسته‌ی محاسبه شده را به دست می‌آوریم.

(اثبات) یک **جواب** با تقریب خوب به کمک تناظر به دست آمده می‌سازیم یا وجود آن را اثبات می‌کنیم.



$$div_k(O) \leq div_k(T) + Mapping(O,T)$$

الگوریتم گنزالز (حریصانه)



الگوریتم

- یک نقطه‌ی شروع دلخواه انتخاب کن.
- نقطه‌ای که بیشترین فاصله را تا نقاط انتخاب شده دارد به مجموعه اضافه کن.
- گام قبل را تکرار کن تا k نقطه به دست بیاید.

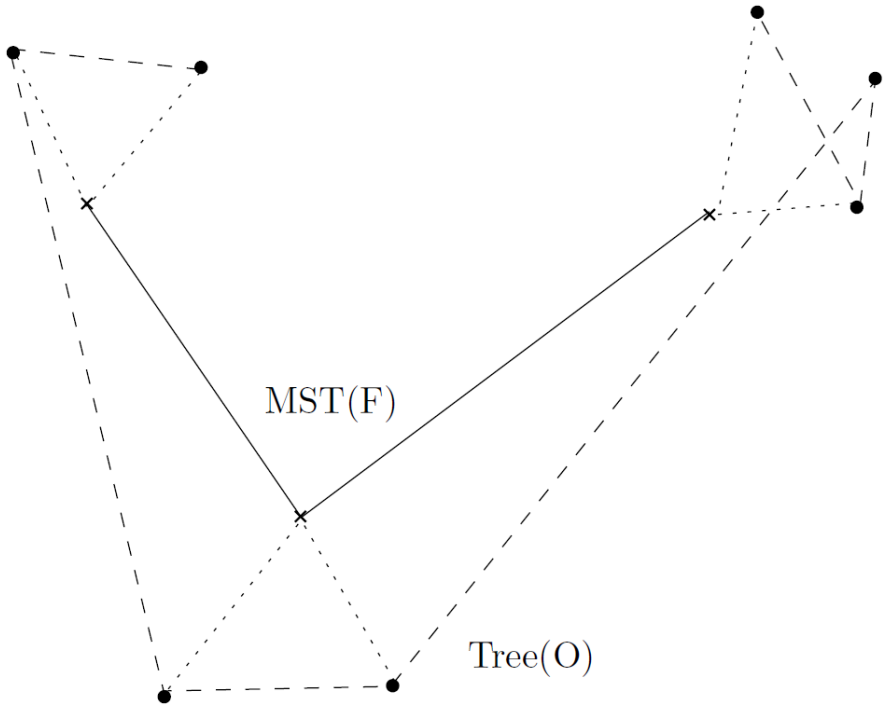
زمان $O(kn)$ و حافظه $O(kn)$

ویژگی ضدپوششی:

- تنوع یال بافاصله برای این مجموعه جواب (Q) را r در نظر بگیرید. در این صورت دو ویژگی زیر برقرار هستند:
  - فاصله‌ی هر نقطه خارج مجموعه Q تا نزدیک‌ترین نقطه‌ی Q حداکثر r است. (تمام نقاط با این شعاع پوشیده می‌شوند.)
  - فاصله‌ی هر دو نقطه‌ی Q از یک‌دیگر حداقل r است. (طبق انتخاب حریصانه‌ی الگوریتم)

بهبود تقریب جواب‌های الگوریتم گنزالز

درخت بافاصله و دور بافاصله



کلیات اثبات

هر نقطه‌ی جواب بهینه را به **نزدیک‌ترین نقطه‌ی** آن در مجموعه‌ی هسته تصویر می‌کنیم.

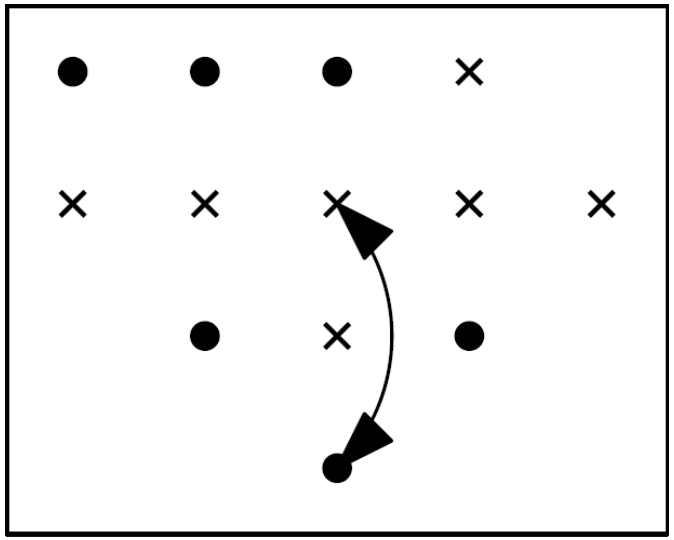
با الگو قرار دادن جواب الگوریتم، یک جواب روی نقاط بهینه می‌سازیم.

طبق خاصیت ضدپوششی می‌دانیم همواره یک جواب هست که همه‌ی یالهای آن حداقل r هستند.

اگر نقاط کمتر از مقدار مورد نیاز بود نقطه‌ی دلخواه اضافه می‌کنیم.

در اثبات t–درخت و t–دور، نقاط تکی هزینه‌ی صفر دارند.

الگوریتم جستجوی محلی



$\bullet \in S \setminus c(S)$

$\times \in c(S)$

ابتدا دورترین دو نقطه را پیدا کن و با k-2 نقطه‌ی دلخواه دیگر به عنوان جواب اولیه در نظر بگیر.

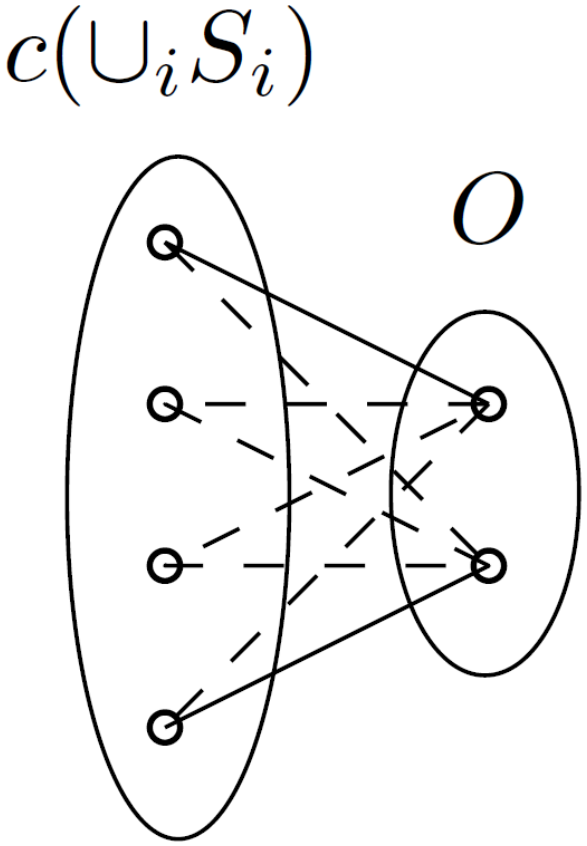
سپس هر بار یک نقطه داخل جواب را با یک نقطه خارج از جواب جابه‌جا کن اگر تنوع جواب حداقل ۱+اپسیلون برابر شود.

زمان $O(kn/\varepsilon \log k)$

ویژگی (نتیجه‌ی شرط خاتمه‌ی الگوریتم)

جمع یالهای مجاور یک رأس از جواب بهینه که در مجموعه‌ی هسته نیست، حداکثر $kr(1+\varepsilon)$ است.

بهبود تقریب خوشه با جستجوی محلی



پیدا کردن تطابق

- تطابق هر رأس اشتراک به خودش
- ساخت گراف G_x و پیدا کردن تطابق کمینه
- ساخت گراف G_y و تکمیل تطابق کمینه

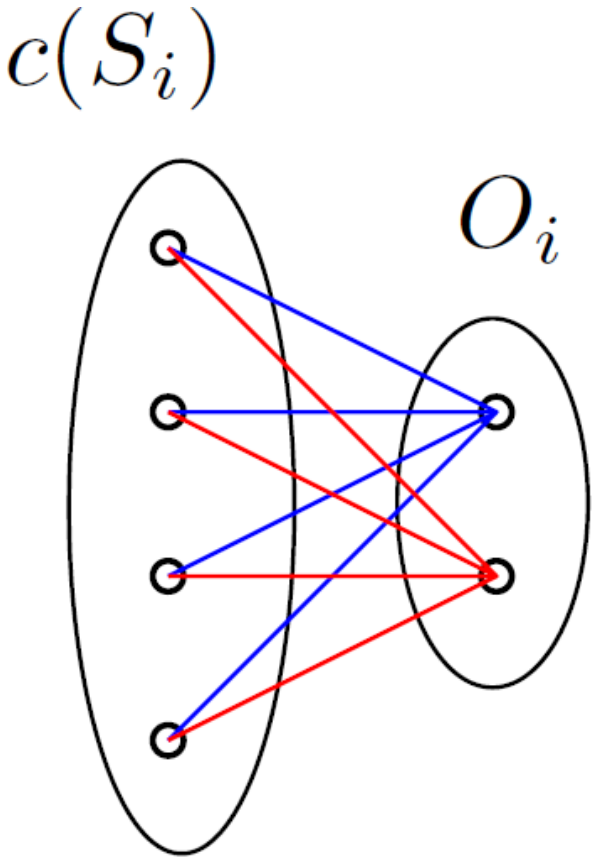
محاسبه‌ی وزن تطابق

- محدودیت‌هایی مثل: طول بزرگترین یال تناظر نمی‌تواند از جواب بهینه بیشتر باشد.
- مقادیر x و y را طوری انتخاب می‌کنیم که $y>x$ باشد و محدودیت‌ها برقرار باشد.

محاسبه‌ی جواب

- با استفاده از نامساوی مثلث جواب بهینه را بر حسب وزن یک خوشه روی نقاط متناظر و وزن تطابق می‌نویسیم.

بهبود تقریب خوشه با ج.م. (افرازی)



محاسبه‌ی تطابق با کمترین وزن (نامساوی شرط خاتمه)

نقاط اشتراک به خودشان تصویر می‌شوند.

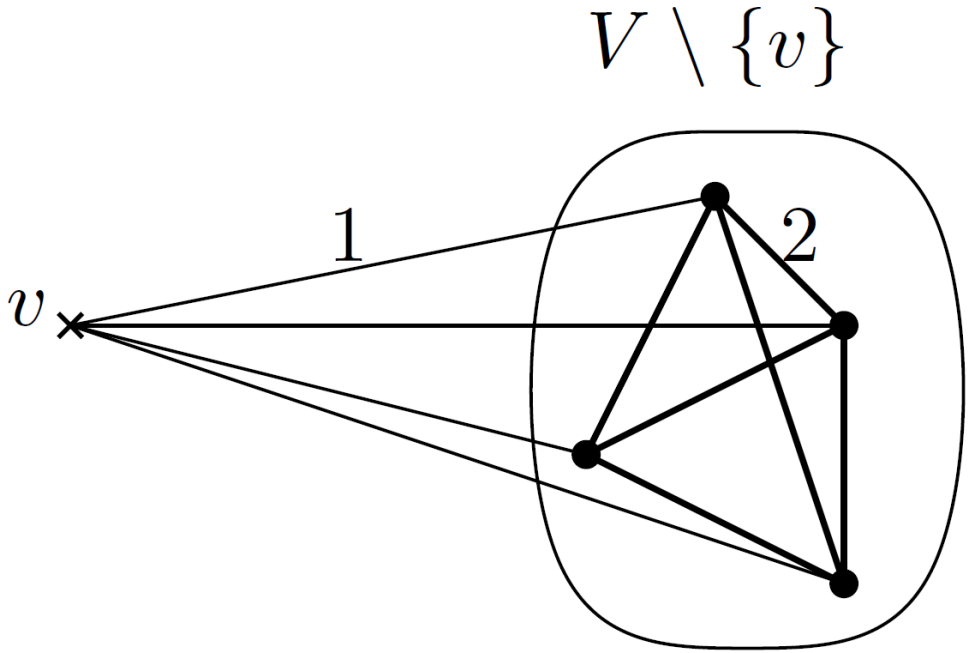
پس از حذف نقاط اشتراک از مجموعه‌ی هسته (نه جواب بهینه)، به یک گراف دوبخشی کامل می‌رسیم.

اثبات شمارشی: جمع یالهای یک گراف دوبخشی کامل و تعداد تطابق‌های آن را حساب می‌کنیم و میانگین را به دست می‌آوریم. کمینه همیشه از میانگین کمتر است.

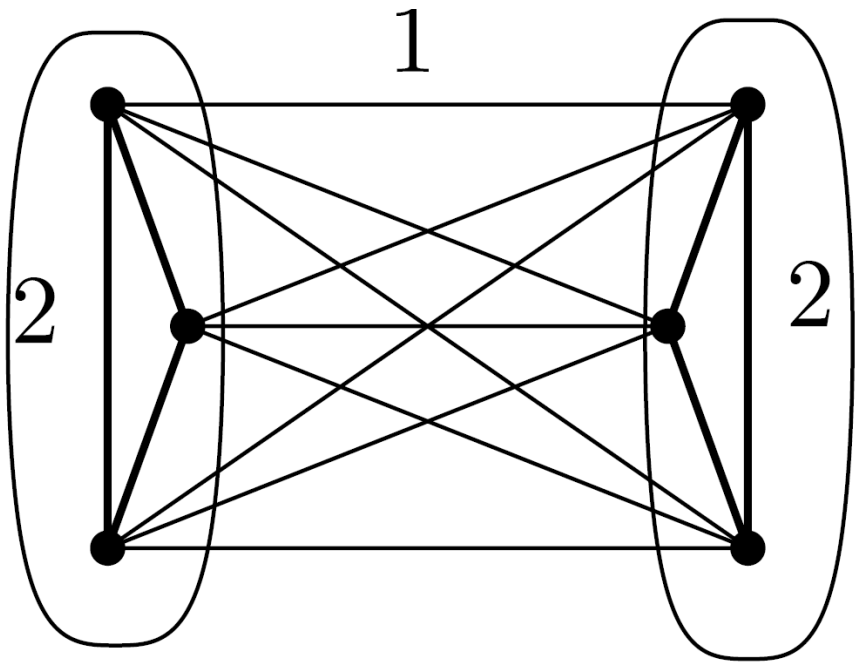
با استفاده از این تطابق و تناظر اشتراکها به خودشان، ممکن است دو نقطه (یکی در اشتراک و دیگری نسبت داده شده به یک نقطه‌ی اشتراک) به یک نقطه (نقطه‌ی اشتراک) تصویر شوند.

بهبود تقریب افراز دوتایی و ستاره

star < 2 clique/k
clique < (k-1) star



clique = A+B+M
(k-1).sum(a) ≥ 2 B



مقایسه‌ی ضریب تقریب‌های قبلی و جدید

ملاک تنوع	قبلی	جدید	جدید افرازی
یال	۳	**کران پایین ۳**	**کران پایین ۳**
خوشه	**۵۱**	$۷ + ۴\sqrt{۲} + \varepsilon \approx ۱۳$	**۶**
ستاره	۱۰۲	$۱۴ + ۸\sqrt{۲} + \varepsilon \approx ۲۶$	۱۲
افراز دوتایی	۲۵۵	$۲۱ + ۱۲\sqrt{۲} + \varepsilon \approx ۳۸$	۱۸
درخت	۶	۴	۴
t-درخت	۶	۴	۴
دور	**۱۲**	**۳**	**۳**
t-دور	۱۲	۵	۵

کارهای آینده

مجموعه‌های هسته‌ی ترکیب‌شونده برای مسایلی که برای آنها مجموعه‌ی هسته‌ی غیر برخط وجود دارد.

بررسی نسخه‌ی پویای مسئله (اضافه کردن امکان حذف)

بررسی نسخه‌ی متحرک مسئله (اضافه کردن امکان جابه‌جا شدن نقاط)

ساخت تصادفی مجموعه‌های اولیه به جای دلخواه بودن آنها در مجموعه‌های هسته‌ی ترکیب‌شونده

افزایش تعداد تکرارهای دیدن ورودی

بهبود ضریب تقریب مسایل گفته شده (مثلاً با تغییر کران تقریب یا اثبات بدون استفاده از مسایل تنوع دیگر)

پیدا کردن الگوریتم‌های دیگر برای حل این مسایل

استفاده از این مسایل برای حل مسایل مشابه مانند مسایل خوشه‌بندی و عکس تنوع

سوالات؟