Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Profesores: José Correa, José Verschae
Prof. Auxiliar: Alberto Vera Azócar

Fundamentos de Redes Sociales
Clustering, Comunidades y Pequeño Mundo

11 de diciembre de 2013

Problema 1 [Comunidades].- Para una campaña de marketing en una red social se ha elegido como target a la "mejor" comunidad (en el sentido de densidad). Para encontrar dicha comunidad considere el Algoritmo 1.

1. Dé el orden del Algoritmo 1
2. Pruebe que el Algoritmo 1 entrega una $\frac{1}{2}$-aproximación de la mejor comunidad, para eso se recomienda
   a) Muestre que si no se tiene la propiedad y denotamos por $S^\star$ a la mejor comunidad, entonces $den(S^\star) > d - 1$
   b) Muestre que si se cumple lo anterior, entonces $\forall v \in S\ deg_S(v) > d - 1$. Concluya.
3. ¿Es posible que en alguna iteración G_d tenga menos densidad que G_{d-1}?

Algoritmo 1 Rank Subgraph en $G = (V, E)$ no dirigido

```
d ← 0, G_0 ← G
while G_d ≠ ∅ do
    d ← d + 1
    G_d ← G_{d−1}
    Remover de G_d todos los vértices con grado menor a d y sus respectivos arcos
end while
Retornar G_{d−1}
```

Problema 2 [Pequeño Mundo].- Según el experimento de Milgram de los seis grados de separación, la mayoría de la gente basaba su decisión de rutear el mensaje en base a sólo dos aspectos: geográfico y ocupacional. Por esto nace la idea de estudiar un modelo como sigue.
Considere un árbol binario $T = (V, E)$, $k = c \log n$ un parámetro que depende de c y $\alpha \geq 0$ dado. Por cada nodo $v \in V$ se sortean k vecinos de largo alcance, con la ley $\mathbb{P}(v \to u)$ proporcional a $2^{-\alpha d_T(v,u)}$.
A diferencia del modelo de Kleinberg, una vez que los links largos han sido sorteados se borran los links de E. Asumiremos que el mensaje parte del nodo v raíz y que existe un camino hacia el target t.
Se desea estudiar el ruteo descentralizado en el nuevo grafo.

1. Muestre que con $\alpha = 1$ y c apropiado el Algoritmo 2 está bien definido w.h.p. y logra entregar un mensaje en tiempo $\mathcal{O}(\log n)$

2. Muestre que para $\alpha < 1$ no hay algoritmo descentralizado que entregue el mensaje en tiempo $\mathcal{O}(\text{poly}(\log n))$. Para esto sea $\beta \in (0, 1-\alpha)$ y considere el sub-árbol $T^\star$ que contiene a t y es de tamaño n^β

3. Muestre que para $\alpha > 1$ no hay algoritmo descentralizado que entregue el mensaje en tiempo $\mathcal{O}(\text{poly}(\log n))$

Nota: Al igual que en el modelo de Kleinberg obtenemos una dicotomía en base al parámetro. Además, así como el modelo de la grilla se puede extender a d dimensiones, este modelo se puede extender a un árbol $b - ario$.

Algoritmo 2 Last Common Ancestor Routing en el árbol T

Sea v el nodo que tiene el mensaje, t el target.
while $v \neq t$ **do**
  Sea u el segundo nodo en el único (v,t)-camino
  $\tilde{T} \leftarrow$ sub-árbol de T enraizado en u
  $T^\star \leftarrow$ sub-árbol de $\tilde{T}$ de altura $d_T(u,t)$
  Enviar el mensaje al nodo en $T^\star$ que esté más cerca de t
end while

Problema 3 [Clustering de Correlación].- En una red social podemos observar si las personas tienen una característica común (son amigos, les gusta la misma música, etc) o bien si tienen características disímiles.
Sea $G = (V, E)$ un grafo completo con etiquetas $+$ ó $-$ en los arcos. Recordemos que el agreement de una partición está dado como el número de $+$ dentro de los cluster más el número de $-$ que los cruzan.
Queremos encontrar un algoritmo que aproxime el problema de max agreement, denotemos OPT la partición óptima y $|OPT|$ el número de agreements. Finalmente para $V_1, V_2 \subseteq V$ denotamos $\delta^+(V_1, V_2)$ como el número de arcos etiquetados con $+$ que van de V_1 a V_2.

1. Sea $\varepsilon > 0$ dado, $OPT(\varepsilon)$ la partición óptima bajo la restricción que cada cluster no síngleton sea de tamaño mayor a εn, muestre que $|OPT(\varepsilon)| \geq |OPT| - \varepsilon n^2/2$

2. Sean $C_1, \cdots, C_k$ los cluster no síngleton de $OPT(\varepsilon)$ y C_{k+1} la unión de todos los síngleton. Denotemos $s_i = |C_i|$ y $e_{ij} = \delta^+(C_i, C_j)$. Explique por que

$$|OPT(\varepsilon)| = \left(\sum_{i=1}^{k} e_{ii}\right) + \left(\binom{s_{k+1}}{2} - e_{k+1,k+1}\right) + \left(\sum_{i \neq j} (s_i s_j - e_{ij})\right)$$

3. Existe un algoritmo $\mathcal{A}$ tal que, dados valores s_i y e_{ij}, $\mathcal{A}(s_i, e_{ij})$ retorna una aproximación de la mejor partición que satisfaga aquellas restricciones (no pruebe esto, use $\mathcal{A}$ como oráculo).

   Si la aproximación de $\mathcal{A}$ es tal que $OPT_{\mathcal{A}(s_i,e_{ij})} \geq OPT_{s_i,e_{ij}} - \varepsilon n^2/2$, proponga un algoritmo $\mathcal{B}$ para el problema de clustering tal que $OPT_{\mathcal{B}(G)} \geq OPT_G - \varepsilon n^2$.

Nota: Discutamos de como se puede usar el resultado anterior en un grafo G' no completo de la vida real.