UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Politècnica Superior d'Enginyeria de Vilanova i la Geltrú

RECUPERACIÓ DE LA INFORMACIÓ — Control 2

Data: 10 de gener de 2020 — **Temps: 2 hores**

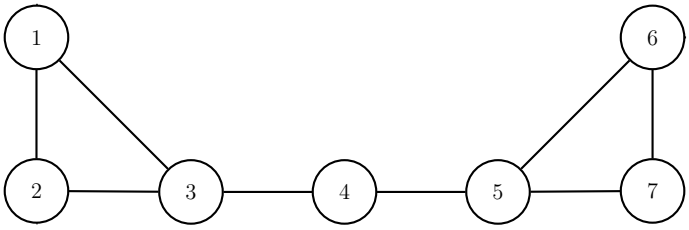
Problema 1 [2,5 punts]

(a) Considera dos nodes d'una xarxa que: i) tenen el mateix grau, ii) un d'ells té un alt coeficient de clustering mentre que l'altre té un baix coeficient de clustering. Mantenint la resta en igualtat de condicions, digues quin dels dos nodes seria l'objectiu si volguessis interrompre la xarxa.

(b) Explica breument què ens indica i per a quina situació ens pot interessar cadascuna de les mesures següents: centralitat de grau, centralitat d'intermediació (*betweenness*), centralitat de proximitat i PageRank.

Problema 2 [2,5 punts]

Donada la xarxa:



(a) Calcula, per cadascun dels nodes de la xarxa, els valors de centralitat de grau, centralitat de proximitat i centralitat d'intermediació (*betweenness*).

(b) Genera una taula de tres columnes, una columna per mesura de centralitat. Introdueix en cada columna les etiquetes dels nodes ordenant-los de més gran a més petit segons els valors calculats en l'apartat anterior. En cas d'empat (dos nodes amb el mateix valor de centralitat per una mesura), fes servir l'ordenació numèrica de l'etiqueta dels nodes.

(c) Obtens el mateix rànquing per les tres mesures? Per què? Quina creus que és la millor?

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Politècnica Superior d'Enginyeria de Vilanova i la Geltrú

Problema 3 [2,5 punts]

Donat un llarg fitxer de text en el qual cada línia conté les dades d'una compra online, volem obtenir el producte que més s'ha venut a cada ciutat. Els camps d'una línia corresponen a la data, hora, ciutat, codi de referència del producte, preu i entitat bancària emissora de la targeta de pagament. Els 6 camps estan separats per ';' com es mostra a continuació:

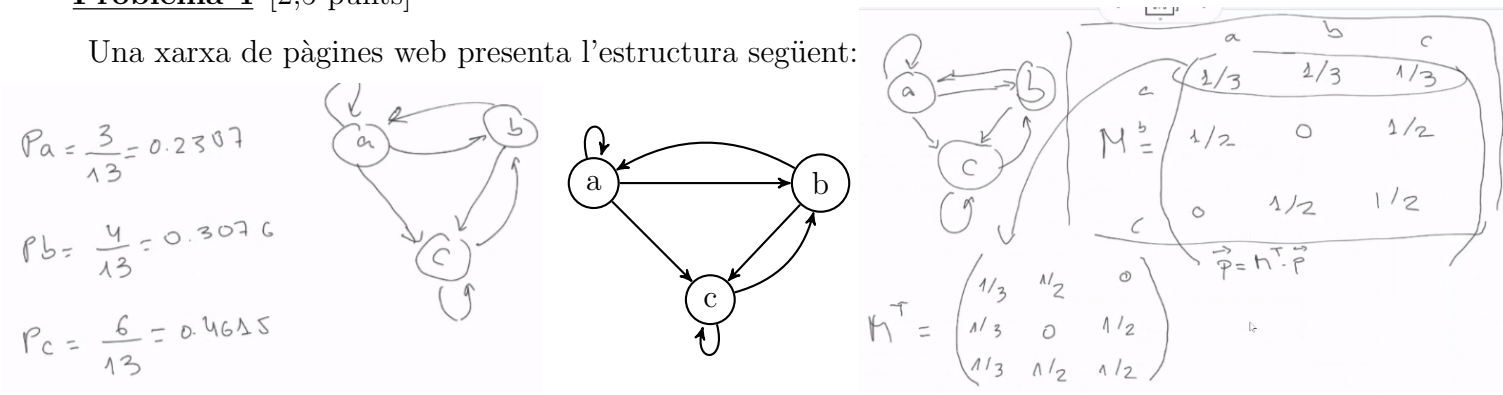
```
2020-01-01;09:00:00;Barcelona;51141001-13;169.00;Kutxabank
```

Descriu com resoldries aquest problema usant el model MapReduce. Pots usar més d'una tasca.

Es valorarà l'eficiència de la solució proposada.

Problema 4 [2,5 punts]

Una xarxa de pàgines web presenta l'estructura següent:



(a) Dona la matriu de probabilitats de transició associada i els valors de PageRank per a totes les pàgines (nodes).

(b) Escriu la matriu de Google usant un factor d'amortiment de 0.8 i el sistema d'equacions resultant per calcular els valors de PageRank de totes les pàgines de la xarxa. Sense resoldre el sistema, indica com varia el PageRank de cada pàgina (més alt o més baix que sense amortiment) justificant la teva resposta.