Nome e Cognome:

Matricola:

Ricerca dell'Informazione nel Web

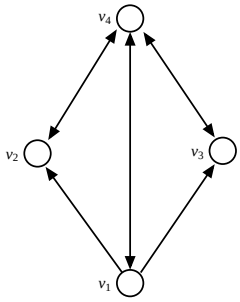
Compito di esame del 9 Settembre 2012, *tempo a disposizione: 90 minuti*
5 punti/problema

Problema 1

1. For which of the following queries are skip pointers most useful, and for which are completely useless? Briefly explain your answers.
   (a) `x AND y`, where `x` is a frequent term and `y` rare
   (b) `x AND y`, where both `x` and `y` are frequent terms.
   (c) `x OR y`, where `x` is a frequent term and `y` rare
   (d) `x OR y`, where both `x` and `y` are frequent terms.
2. Write a pseudocode of an algorithm for merging two postings lists for a query of the type `term1 AND term2` using skip pointers.

Problema 2

1. We are given the following graph. Write down all the necessary equations needed to calculate the pagerank, for a general teleporting probability α.
2. Compute the pagerank of each node for teleporting probability $\alpha = 1/2$.
3. Prove that for any graph the pagerank of each node is at least α/N.



Problema 3

1. Write the *tf*$\times$ *idf* weighting equation. Explain what each term represents, and the reasoning about the equation.
2. Consider an IR system where we use the *tf*$\times$ *idf* weighting scheme. We compare three pairs of documents:
   (a) Two docs that have only frequent words (the, a, an, of, etc.) in common.
   (b) Two docs that have no word in common.
   (c) Two docs that have many rare words in common.

   Rank the above pairs according to their cosine similarity score and explain convincingly the reasoning behind your choice.

3. State two reasons that in IR we usually use cosine similarity instead of Euclidean distance.

Problema 4

1. What are the roles of front queues and back queues in Mercator's crawler URL frontier scheme? Explain briefly how they work.

2. Usually when we start crawling we start with several seed pages. Why is this necessary?