Introduction to Web Science

Assignment 7

PD Dr. Matthias Thimm

thimm@uni-koblenz.de

Ipek Baris Schlicht

ibaris@uni-koblenz.de

Kenneth Skiba

kennethskiba@uni-koblenz.de

Institute of Web Science and Technologies
Department of Computer Science
University of Koblenz-Landau

Submission until: 19.01.2021, CEST 23:59

1 Generative Models for the Web   (25 points)

We provide you a file called `sample_simple_english_wiki.txt`. The file contains a text snippet from Simple English Wikipedia. Your tasks are as follows:

1. Create a probabilistic generative model of text by sampling from the distribution of 1) word lengths 2) frequency of each character in the given text, similar to the one presented in the video slides[1].

   Modelling choices:

   - Your model should generate one word at a time according to the distribution of the word lengths.

   - Within each word, generate characters according to the distribution of character frequencies.

   - Your model should only consider lowercase letters `[a-z]` and numbers `[0-9]`. All uppercase letters should be converted to lowercase. Other characters such as punctuation should be excluded.

   Generate a text of 5,000 words with your model. Please upload your generated text as an individual file, but do not include it in the PDF document.

2. Plot the probability distribution of word lengths of 1) the original text 2) the text you generated in one plot. Save the plot as png file.

3. Discuss the resulted plots and generated text (max half page).

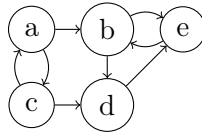For this task, you are allowed to use only string, regex, numpy, matplotlib as library.

---

[1] https://en.wikiversity.org/wiki/Web_Science/Part2:_Emerging_Web_Properties/Generative_Models_for_the_Web/Sampling_from_a_probability_distribution

2 Directed Graphs   (30 points)

Consider the following directed graph G:

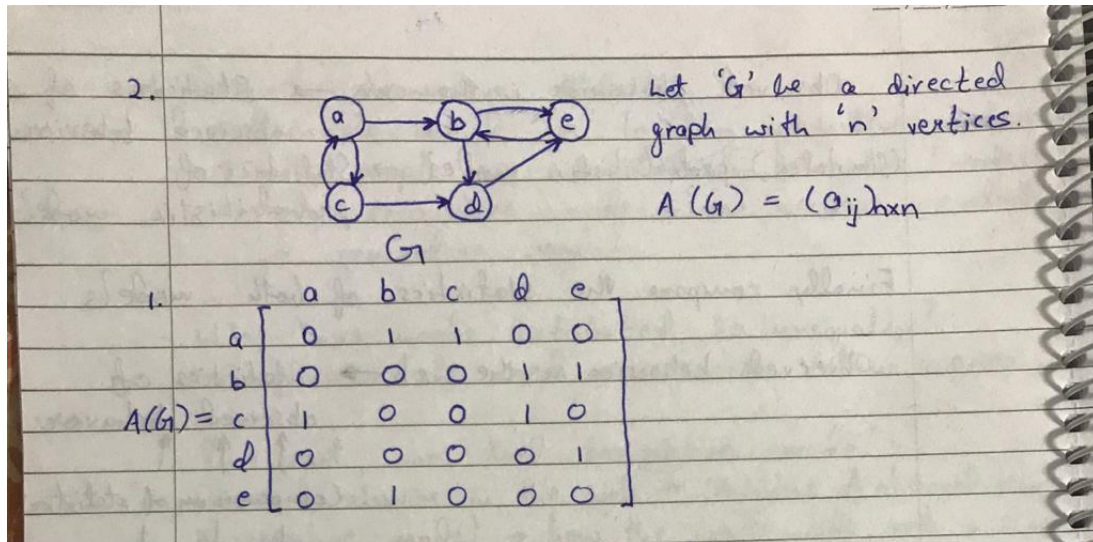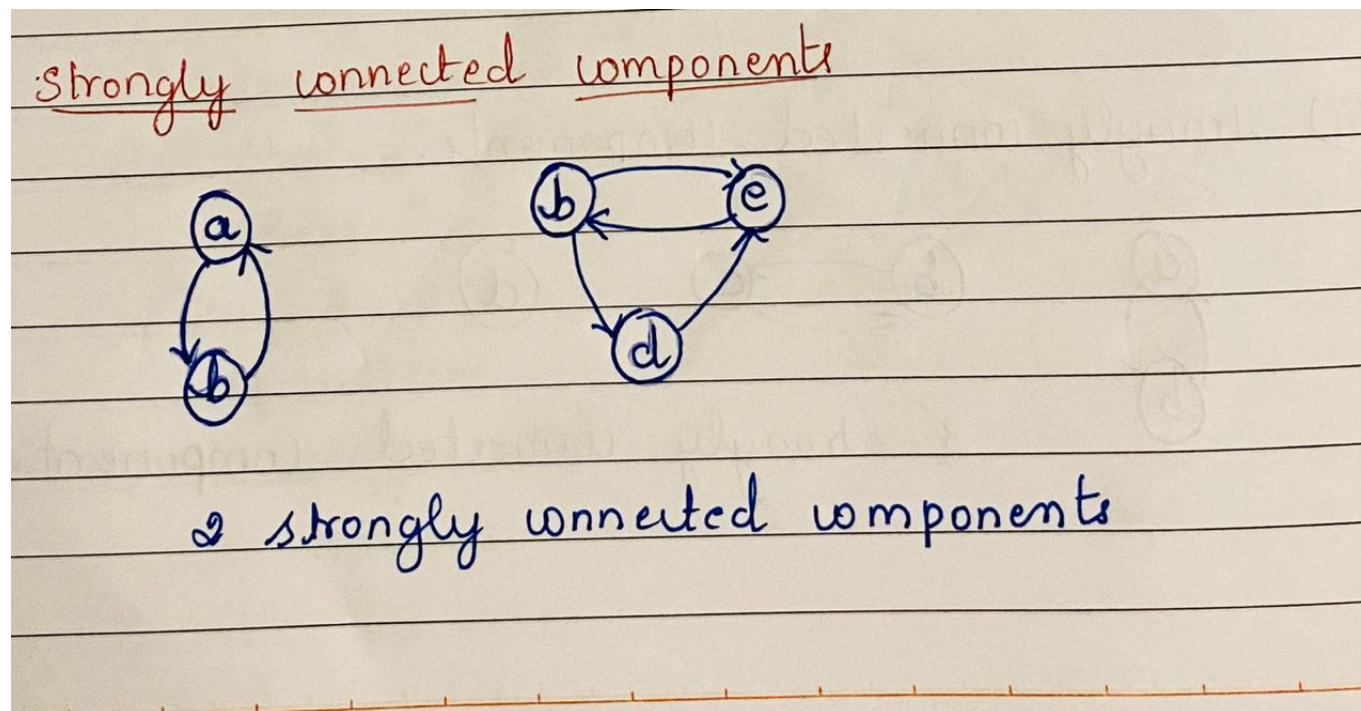
1. Write down the adjacency matrix A of the graph G. adf

Figure 1: Adjacency matrix

2. Calculate the In- and Outdegree of every vertex.

Figure 2: In- and Outdegree

3. Highlight all strongly connected components.

Figure 3: Strongly connected components

4. Construct **one** graph G', which is *isomorphic* to G.

Consider the following graph H.

5. Assuming vertex 11 is our goal vertex. In which order will the nodes be expanded when breadth-first search, and depth-first search is used?
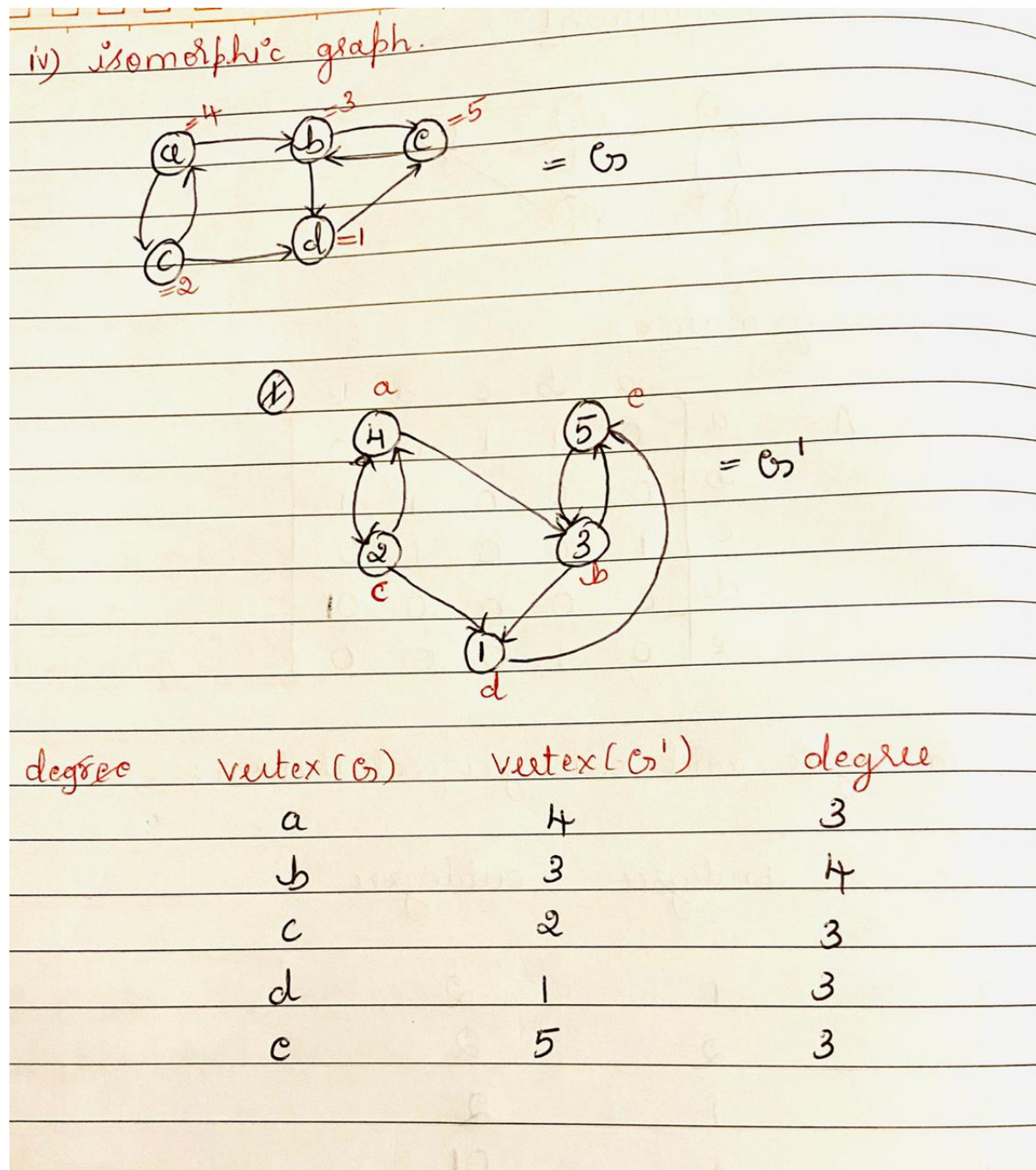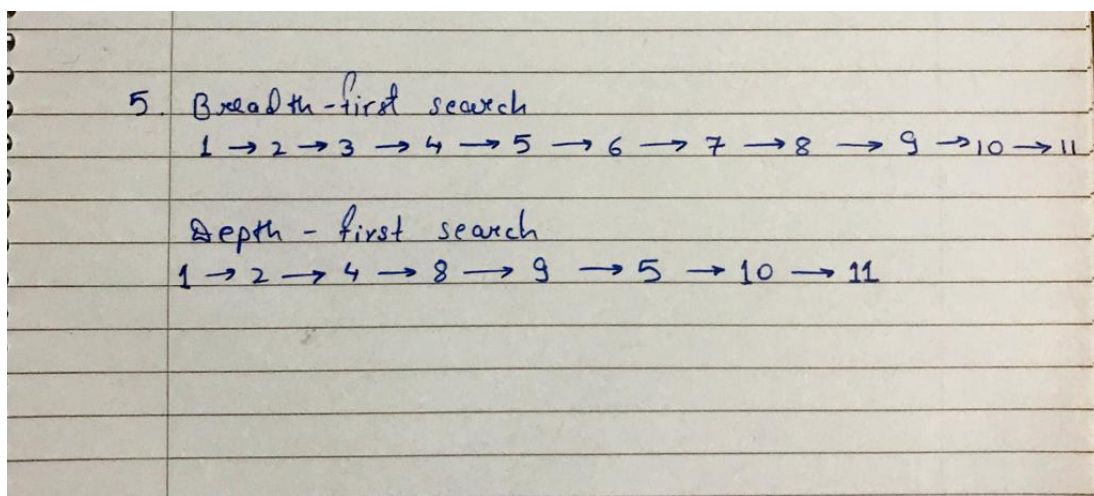
iv) isomorphic graph.

= G

= G'

degree	vertex (G)	vertex (G')	degree
a	4	3
b	3	4
c	2	3
d	1	3
e	5	3

Figure 4: Isomorphic

Figure 5: Breadth-first search, and Depth-first search

3 Undirected Graphs (25 points)

Let $G = (V, E)$ be an *undirected graph.* Using Definitions 1,2,3 prove that Theorem 1 holds.

Definition 1 *The function $e(v)$ of a vertex $v \in V$ returns the longest distance between v any other vertex of G:*

$$e(v) = \max_{w \in V} d(v, w)$$

Definition 2 *The* diameter $diam(G)$ *of G is the greatest $e(v)$ of any vertex in G:*

$$diam(G) = \max_{v \in V} e(v)$$

Definition 3 *The* radius $rad(G)$ *of G is the smallest $e(v)$ of any vertex in G:*

$$rad(G) = \min_{v \in V} e(v)$$

Theorem 1 $rad(G) \leq diam(G) \leq 2 * rad(G)$

3) **undirected graphs**

Theorm 1 : $rad(G) \leq diam(G) \leq 2 * rad(G)$

we have,

$$rad(G) = \min_{v \in V} e(v) \quad\text{------ def (3)}$$

$$diam(G) = \max_{v \in V} e(v) \quad\text{------ def (2)}$$

where

$$e(v) = \max_{w \in V} d(v, w) \quad\text{------ def (1)}$$

$$\therefore rad(G) \leq diam(G) \quad\text{------ ①}$$

Let c be a vertex at a minimal distance from all the vertices.

$$d(c, v) \leq rad(G)$$
$$d(c, u) \leq rad(G)$$

where $u, v \in V(G)$ and d is the distance using triangular inequality,

$$d(u, v) \leq 2 \, rad(G)$$

$$diam(G) \leq 2 \, rad(G) \quad\text{------ ②}$$

with ① and ② we can proove that
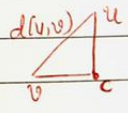$$rad(G) \leq diam(G) \leq 2 \, rad(G)$$

Figure 6: Undirected Graphs