15-826
Multimedia Databases and Data Mining
Homework 2 Solutions

Ina Fiterau
mfiterau@cs.cmu.edu

December 6, 2011

1 Fractals

1. See source code in *Q1. Fractals/Battlements.py*.

2. See plots of B_2 and B_3 in Figure 1.

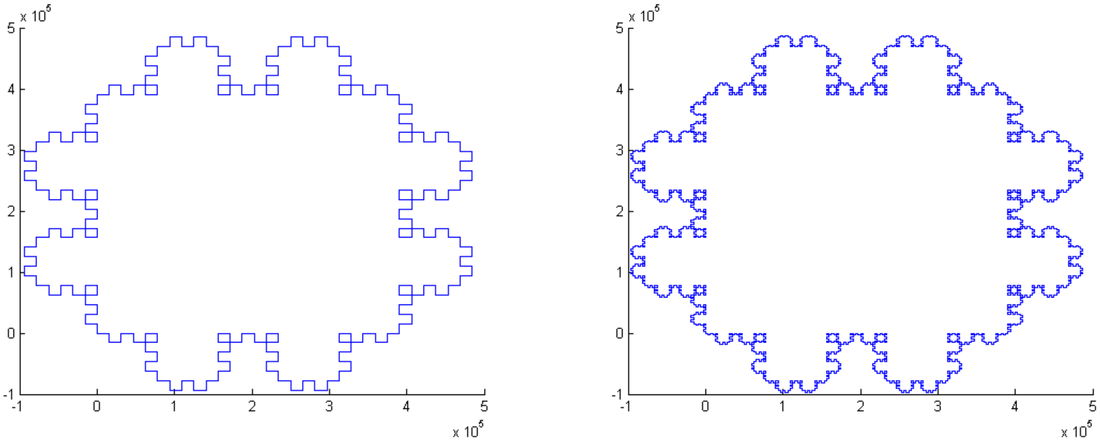

Figure 1: B_2 (left) and B_3 (right)

3. Fractal dimension: $\frac{\log 9}{\log 5} \approx 1.37$

4. Hausdorff plot of the corner points is in Figure 2. It was obtained by runnung fdnq on the corner points of the curve and then running the boxcounting script (also in the fdnq archive) on the resulting *.points* file.

5. The Hausdorff dimension of the B_4 curve is also 1.37. Note that the dimension of the cloud of corner points can be obtained from the Hausdorff plot by negating the slope of the steadily decreasing portion: $D = -\frac{\log N(r_1) - \log N(r_2)}{\log r_1 - \log r_2}$. The minus sign comes from the dependence $N = Cr^{-D}$.

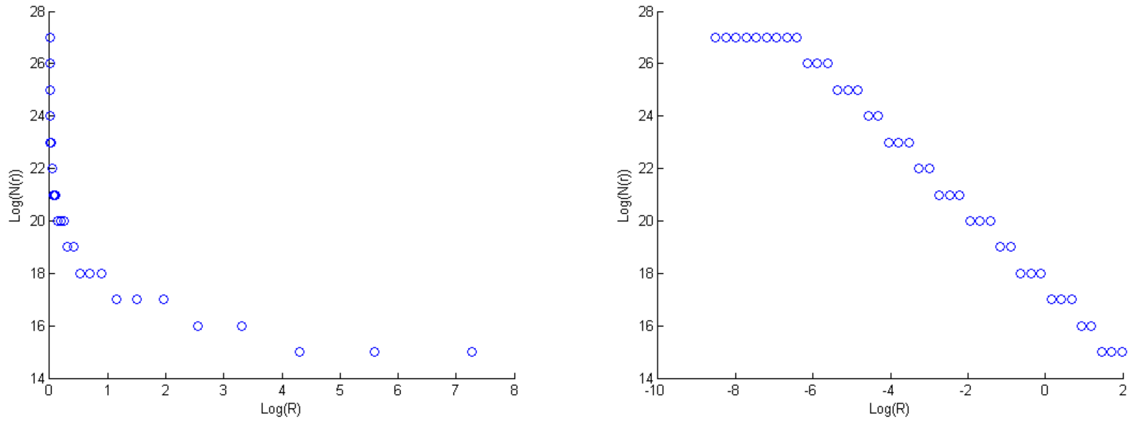

Figure 2: The Hausdorff plot. r vs logN (left) and logr vs logN (right)

2 Multifractals (Bursty Time Series)

1. See code in *Q2. Multifractals/BurstySeries.py*

2. The plots for b=0.5,b=0.7 and b=0.9 are show in Figure 3.

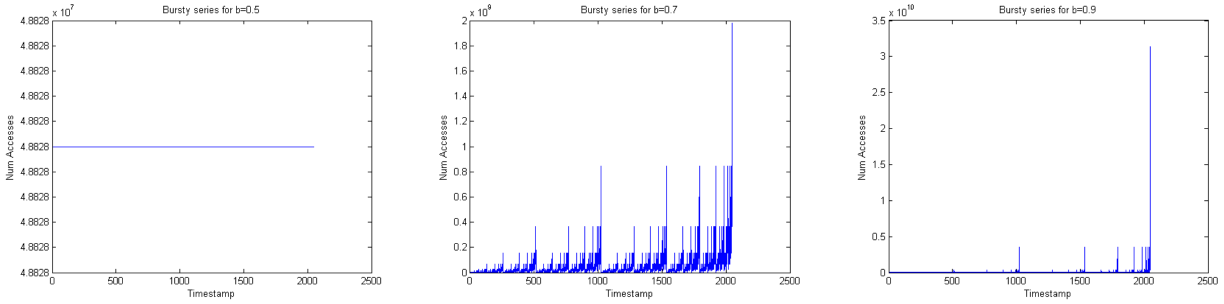

Figure 3: Bursty time series. b=0.5 (left); b=0.7 (center); b=0.9 (right)

3. Correlation integral plots (obtained by running FDNQ) are shown in 4

4. To obtain the bias estimator, solve the quadratic equation $2b^2 - 2b + 1 - 2^{-D} = 0$

$$b = \frac{2 + \sqrt{4 - 2(1 - 2^{-D})}}{4}$$

.

$$\begin{aligned} d = 0.99 &\quad \rightarrow \quad b = 0.53 \\ d = 0.80 &\quad \rightarrow \quad b = 0.68 \\ d = 0.35 &\quad \rightarrow \quad b = 0.87 \end{aligned}$$

NOTE: Approximate answers are ok for this question.

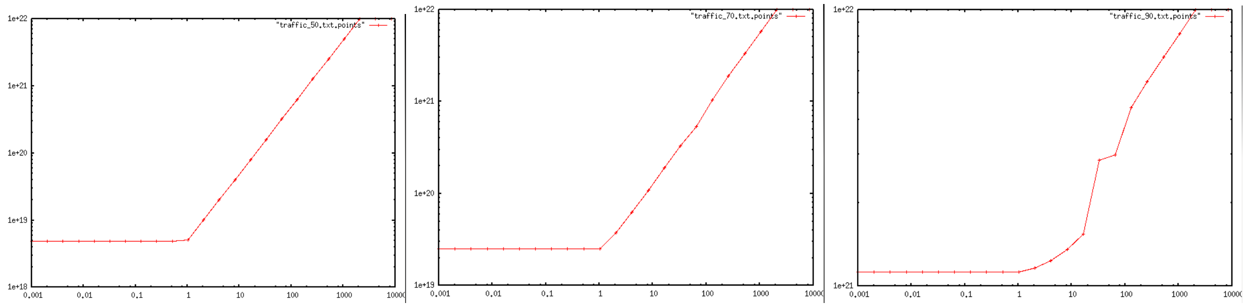

Figure 4: Correlation Integrals. b=0.5 (left); b=0.7 (center); b=0.9 (right)

3 Power laws

The answers to this question refer to the included *Q3. Power laws/stars.csv*. Due to issues with the version of the dataset that is on the website, full points were given for the first two items of this question if an FDNQ plot was provided.

1. The correlation integral plot is shown in Figure 5

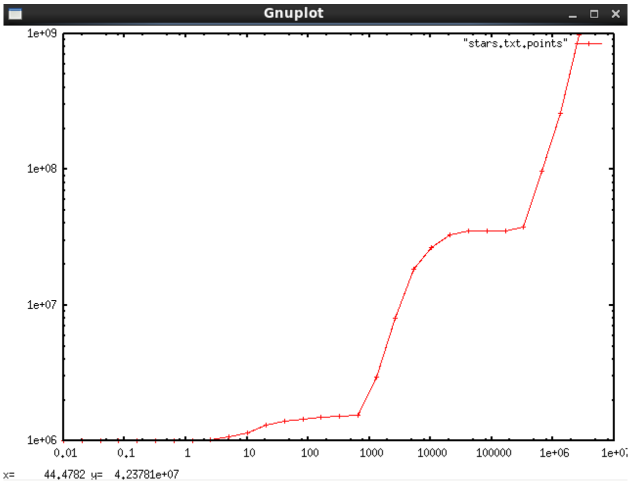

Figure 5: Correlation Integral

2. The data points are distributed in clusters of clusters. The point clusters have radius r=10 and are 1000 units apart. These small clusters are grouped into large clusters of radius R=5000 that are $5 * 10^5$ units apart.

   The dataset on the website, which contains real data, has no underlying structure.

   The file *Q3. Power laws/stars.csv* shows how the cluster data was generated.

4 String Editing Distance

1. See code in *Q4. String Editing Distance/sed.py*.

2. The replacements are as follows:

```
Replacements for  soeaking
['peking', 'seedling', 'sterling']
Replacements for  prohect
['protect', 'prophecy', 'project']
Replacements for  sholder
['shoulder', 'smolder', 'solder']
Replacements for  laptpp
['layton', 'layup', 'lipton']
Replacements for  folowing
['fleeing', 'folksong', 'filipino']
Replacements for  beowser
['bender', 'beset', 'border']
Replacements for  paccing
['puccini', 'bagging', 'batwing']
```