Netzwerkanalyse Projektdokumentation

30.09.2016

Viktor Dinkel 4454398

1. Projektbeschreibung
   In dieser Dokumentation geht es um die Generierung, Visualisierung und Analyse eines Netzwerks aus Börsendaten mit der Anleitung gegeben im Paper „Network Analysis of the Stock Market“. Mit dieser Netzwerkanalyse wird ein Portfolio von verschiedenen Aktien zusammengestellt, deren Kurse korrelieren und sie mit einer Zentralitätsfunktion eine möglichst optimierte Auswahl darstellen. Das beinhaltet, dass sie sowohl den Markt repräsentieren, als auch auf verschiedene Module aufgeteilt sind, um das Risiko zu minimieren.

2. Projektziele
   1. Beschaffung von Börsendaten
   2. Generierung eines Korrelationsnetzwerks
   3. Module innerhalb des Netzwerks identifizieren und visualisieren
   4. Portfolio mittels Zentralitätsfunktion zusammenstellen
   5. Auswertung der Module und des Portfolios

3. Durchführung
   3.1. Datenbeschaffung
      3.1.1. Python *Yahoo Query Language (YQL)*
      Die Daten werden, im Gegensatz zum Paper, nicht vom *Center for Research in Security Prices* (CSRP) heruntergeladen, da es eine kostenpflichtige Plattform ist. Stattdessen Wird die kostenfreie Quelle für historische Börsendaten auf *Yahoo! Finance* genutzt. Ein Python-Script stellt die Anfrage per URL an die API, um die gewünschten Börsendaten innerhalb eines Zeitintervalls herunterzuladen und in einer lokalen Datei zu speichern. Die genutzte Schnittstelle heißt Yahoo Query Language (YQL, https://developer.yahoo.com/yql/) und folgt einer SQL-ähnlichen Syntax. Dafür wurde zunächst eine Liste aller Währungen extrahiert (http://www.nasdaq.com/screening/companies-by-industry.aspx?exchange=NASDAQ), für die jeweils mit dieser Anfrage alle Kursdaten für das Jahr 2015 ermittelt wurden.

      ```
      SELECT * from yahoo.finance.historicaldata
      WHERE symbol = "CFO"
      AND startDate = "2015-01-01"
      AND endDate = "2016-01-01"
      ```

      3.1.2. Dateneigenschaften
      Die Antwort dieser Anfrage ist im JSON-Format und beinhaltet u.a. Angaben zu Tageskursen und dem verfügbaren Tagesvolumen. Zu den insgesamt 2614 Unternehmen existieren jeweils bis zu 252 Kurseinträge innerhalb des Jahres 2015. Es sind allerdings nicht alle heruntergeladenen Datensätze vollständig, weshalb sie in der fortgeschrittenen Durchführung aus der Analyse ausgeschlossen werden.

3.2. Netzwerkgenerierung

3.2.1. Matrix *M* aus logarithmischen Kursänderungen

Gegeben ist nun eine Datenmatrix aus Kursdaten, jede Zeile enthält eine Zeitreihe von Schlusskursen einer Aktie über den Zeitraum von einem Jahr (2015). Das sind die Aktienwerte zum Ende eines Tages. Die absoluten Kurswerte sind jedoch weniger interessant, stattdessen werden daraus Kursänderungen generiert mit der Formel (2).

$$r_i(t) = \ln\left[\frac{p_{c,i}(t)}{p_{c,i}(t-1)}\right]. \qquad (2)$$

Das ist der Logarithmus aus dem Schlusskurs p_c einer Aktie i zum Zeitpunkt t geteilt durch ihren Schlusskurs am Tag davor. Somit werden die absoluten Schlusskurse zu logarithmischen Änderungen des Schlusskurses transformiert. Dadurch lassen sich korrelierende Kursänderungen besser identifizieren.

3.2.2. Korrelationsmatrix *A'*

Um die Adjazenzmatrix *A* zu erstellen, müssen die Kursänderungen jeweils verglichen werden. Somit wird zunächst eine Matrix *A'* erstellt, die ein Ähnlichkeitsmaß [0,1] zwischen den einzelnen Aktienkursen darstellt und $A'_{ii} = 0$. Dafür wird die Matrix *M* mit der Formel (1) zur Korrelationsmatrix *A'* transformiert.

$$c_{ij} = \frac{\sum_t \left[\left(x_i(t) - \bar{x_i}\right)\left(x_j(t) - \bar{x_j}\right)\right]}{\sqrt{\sum_t \left(x_i(t) - \bar{x_i}\right)^2}\sqrt{\sum_t \left(x_j(t) - \bar{x_j}\right)^2}}, \qquad (1)$$

c_{ij} ist der Eintrag A'_{ij} in der Korrelationsmatrix. Das $x_i(t)$ und $x_j(t)$ sind die berechneten Kursänderungen aus 3.2.1 von Aktien *i* und *j* zum Zeitpunkt *t*. Das $\bar{x}$ ist der jeweilige Durchschnittswert aller Kursänderungen einer Aktie.

3.2.3. Adjazenzmatrix *A*

Nun muss entschieden werden, bei welcher Ähnlichkeit bzw. Korrelationsstärke eine Kante zwischen zwei Knoten generiert werden soll. Genauer gesagt, ab welchem Schwellenwert θ in *A'* der entsprechende Wert der Adjazenzmatrix *A* eine 1 bekommt.

$$A_{ij} = \begin{cases} 1 & \text{if } c_{ij} >= \theta \text{ and } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

Somit ist θ der direkte Parameter für die Netzwerkdichte. Je kleiner θ ist, desto dichter sind die Kanten im Netzwerk. Um das richtige θ zu identifizieren, wurden mehrere θ ausprobiert (Abbildung 1).

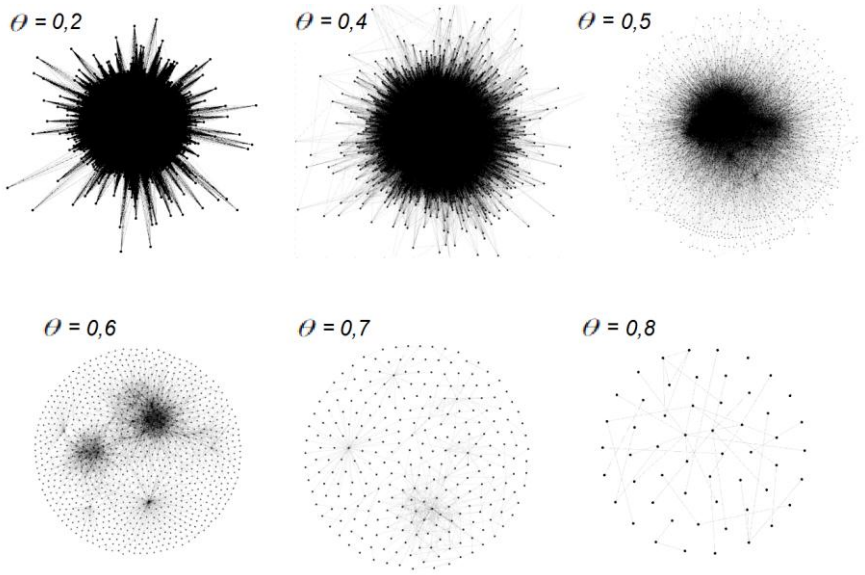

Abbildung1: Visualisierung des Netzwerks mit Schwellenwert θ = [0.2,0.4,0.5,0.6,0.7,0.8]. Das angewandte Layout ist der Fruchterman-Reingold Algorithmus

Die entsprechenden Netzwerke wurden ausgewertet und es hat sich herausgestellt, dass bei θ = 0,6 die Netzwerkeigenschaften am interessantesten sind (Abbildung 2). Die Kantendichte ist gerade richtig, so dass sich Cluster visuell differenzieren lassen. Bei kleinerem Schwellenwert bildet das Netzwerk einen großen Ball, wohingegen bei größerem Schwellenwert das Netzwerk eine zu geringe Kantendichte besitzt.

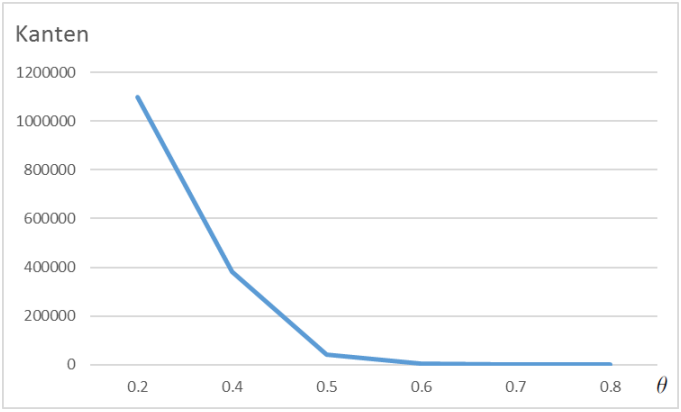

Abbildung2: Verhältnis zwischen dem Schwellenwert θ und der Anzahl der Kanten

Zur Gradverteilung sei noch gesagt, dass es einer *power law* folgt, d.h. das generierte Aktiennetzwerk repräsentiert ein Scale-free Netzwerk (Abbildung 3). Somit gibt es wenige Knoten mit sehr großem Grad und viele Knoten mit niedrigem Grad. Der loglog-Plot dieses Knoten-Gradverhältnisses bildet eine gerade Linie, was die *power law* visualisiert.

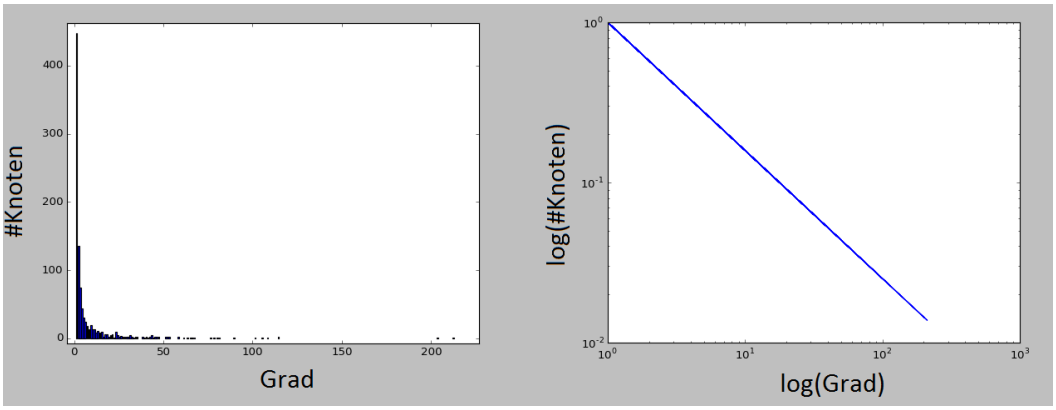

Abbildung 3: Verteilung des Knotengrads und der Anzahl entsprechender Knoten (links) und der loglog-Plot davon (rechts)

3.3. Identifizierung und Visualisierung von Modulen

3.3.1. Fruchterman-Reingold Layout versus Force Atlas

Das Layout in der Abbildung 1 wurde mit dem Fruchterman-Reingold Algorithmus in Gephi definiert. Dafür wurde eine Exportfunktion in Python geschrieben, so dass die Adjazenzmatrix *A* im Gephi-Kompatiblen Format in eine .CSV-Datei geschrieben wird. Nach dem Import in Gephi, kann Fruchterman-Reingold als Layout-Algorithmus ausgewählt werden. Dadurch können einzelne Cluster mit verschiedenen Größen visuell identifiziert werden. Die Anordnung im Kreis ist allerdings zur visuellen Unterscheidung der Module nicht optimal, weshalb das vom Paper abweichende Layout Force Atlas gewählt wurde (Abbildung 4a).

3.3.2. Modularität

Um die Module innerhalb des Netzwerks zu erkennen, wurde die Gephi-Funktion *Modularität* verwendet. Als Parameter wurde die Auflösung 5.0 verwendet, da eine höhere Auflösung als 1.0 die Anzahl der Gemeinschaften reduziert, es also weniger, dafür aber Größere, Gemeinschaften berechnet. Das Resultat davon (Abbildung 4b) erkennt die Module sehr gut, selbst das kleine Modul im rechten Bereich, welches mit der bräunlichen Farbe markiert wurde, wird sich in der weiteren Analyse legitimieren. Dafür wurde folgende Farbzuordnung gewählt:

Modul 0: Grün, Modul 5: Braun, Modul 7: Pink, Modul 9: Violett

Abbildung 4a: Force Atlas Layout. Die Knotengröße skaliert ein wenig mit dem Knotengrad

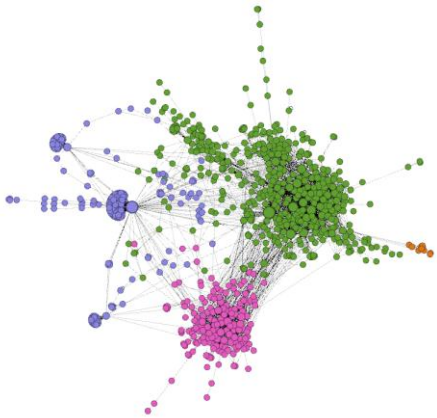

Abbildung 4b: Vier farblich markierte Modularitätsklassen

3.3.3. Sektoranalyse

Nachdem nun das Netzwerk visualisiert und die Module identifiziert wurden, können diese Informationen aus dem Datenlabor in Gephi als .CSV exportiert und in das Pythonscript zur weiteren Auswertung importiert werden. Somit können dort die Modulinformationen sowie weitere Metadaten, wie der Zugehörigkeit einer Aktie zum entsprechenden Sektor (Finanzen, Technologie, etc.), in Verbindung gebracht und statistisch ausgewertet werden (Abbildung 5). Es zeigt sich, dass das violette und grüne Modul aus Aktien verschiedener Sektoren zusammengesetzt ist und kein Sektor dominiert. Im Gegensatz dazu ist das pinke Modul sehr vom Finanzsektor geprägt und das braune Modul vom Energiesektor. Diese Erkenntnis steht im Einklang mit dem Resultat des Papers, in dem ebenfalls festgestellt wurde, dass sowohl heterogene, als auch homogene, Module identifiziert wurden.

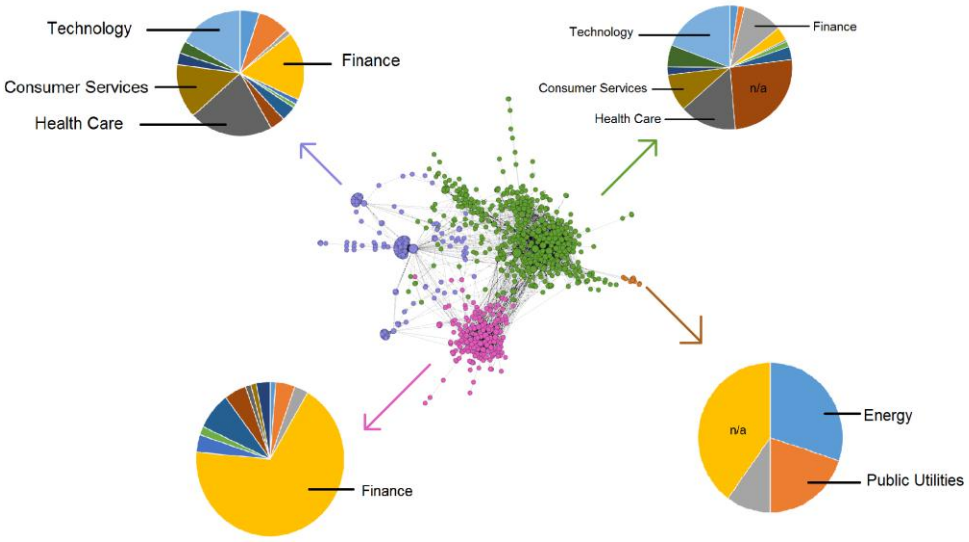

Abbildung 5: Aufschlüsselung der Module in die Anteile von Sektoren der Aktien

3.3.4. Modulrepräsentanz des Marktes

Zunächst wird die Leistung $W_{i,t}$ einer Aktie i zum Zeitpunkt t definiert als die Differenz des Aktienwertes V_i zum Zeitpunkt t und zum Anfangszeitpunkt t_s, das ganze dividiert wieder durch den Aktienwert $V_{i,ts}$ am Anfang der untersuchten Zeitperiode. Der Aktienwert V_i ist das Produkt aus Aktienkurs und dem Aktienvolumen.

$$W_{i,t} = \frac{V_{i,t} - V_{i,t_s}}{V_{i,t_s}} *$$

Wie ist nun die Leistung der Module und wie gut repräsentieren sie den Markt? Um das festzustellen, wird zunächst die Leistung aller Aktien berechnet. Ebenso wird die Leistung der Aktien einzelner Module berechnet, was dann alles zusammen geplottet wird (Abbildung 6). Hier ist zu erkennen, dass die Marktspitzen von unterschiedlichen Modulen beeinflusst werden. Das violette Modul 9 scheint dabei den größten Markteinfluss zu besitzen. Aber auch das vom Finanzsektor definierte Modul 7 hat einen erheblichen Einfluss und trägt ebenfalls erheblich zur Marktdynamik bei. Das größte Modul 0 ist eher träge, hat aber auch in der mittleren Region eine deutliche Spitze. Hier ist nun auch die Abgrenzung des kleinen Moduls 5 vom großen Modul 0 gerechtfertigt, da nun zu sehen ist, dass sich dessen Leistung zum Teil deutlich von der

Leistung des trägen Moduls 0 abhebt. Zusammenfassend sei nun gesagt, dass die Netzwerkgenerierung bis zur Modularisierung den Markt sehr gut in seine Bestandteile aufgeschlüsselt hat. Die Leistungsspitzen können alle fast eindeutig einzelnen Modulen zugeordnet werden.

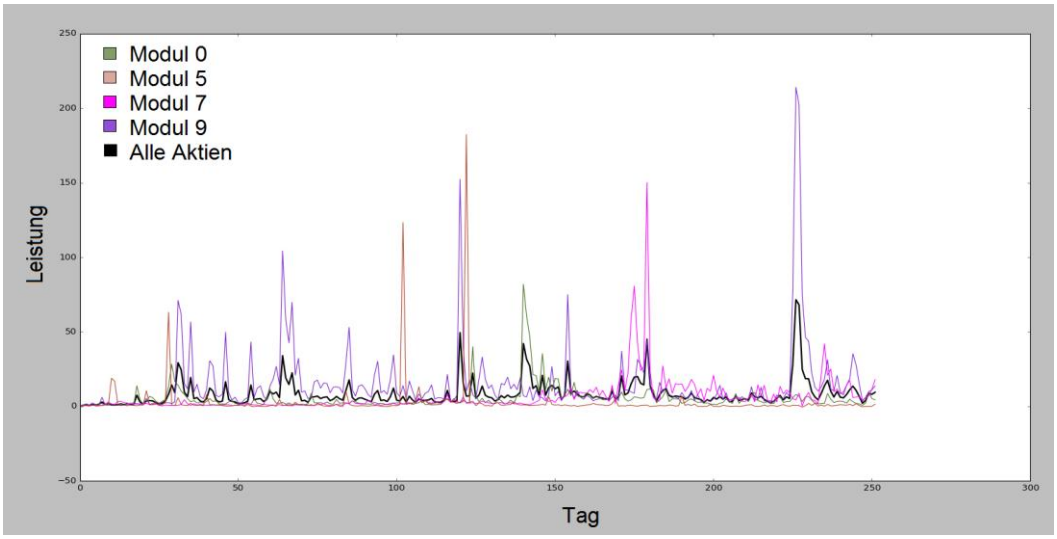

Abbildung 6: Plot der Leistung aller Aktien sowie der Leistung einzelner Module

3.4. Zentralitätsfunktion und Portfolio

3.4.1. Zentralitätsfunktion

Jetzt muss entschieden werden, welche Aktien ein sinnvolles Portfolio ergeben. Nach welchem Maß werden die Kandidaten ausgewählt? Im Paper ist von einer Zentralitätsfunktion die Rede, nach der ein geeignetes Portfolio generiert werden kann. Dabei werden zunächst die Zentralitäten degree- (C_d), closeness- (C_c) und betweenness-centrality (C_b) einzelner Knoten berechnet. Die Zentralitätsfunktion gewichtet dann diese Zentralitäten mit b_1, b_2 und b_3 und errechnet aus der Summe der gewichteten Zentralitäten den Score C_{avg}. Hier sind die Gewichte $b_{1\text{-}3}$ gleichverteilt:

$$C_{avg} = \frac{1}{3}C_d + \frac{1}{3}C_b + \frac{1}{3}C_c \quad *$$

Die Top-X Aktien mit dem höchsten C_{avg} bilden dann das Portfolio. Die genaue Verteilung der Gewichte ist ein Optimierungsproblem und wird in diesem Projekt nur empirisch bestimmt. C_d wird somit mit 0,3 gewichtet, C_b mit 0,2 und C_c mit 0,5. Das hat in der Leistungsauswertung des Portfolios zu einem befriedigenden Ergebnis geführt. In der Praxis würde die Lösung des Optimierungsproblems die optimale Verteilung der Gewichte ergeben, so dass ersichtlich wird, welche Zentralitäten maßgeblich für das Marktverhalten sind. In diesem Projekt hat es sich ergeben, dass die closeness-centrality und die degree-centrality den größten Anteil am Marktverhalten besitzen. Die Bestimmung der Gewichte kann den Schwerpunkt der Portfoliozusammensetzung deutlich beeinflussen. In Abbildung 7 ist zu sehen, wie unterschiedlich der Schwerpunkt des Portfolios, je nach gewählten Gewichten, sein kann.

Abbildung 7: Die Aktien der ausgewählten Portfolios sind dunkelgrau eingefärbt und vergrößert. Rechts ist eine Gleichverteilung Cd = Cb = Cc = 1/3. Links ist die Verteilung Cd = 0.02, Cb = 1/3, Cc = 0.68.

3.4.2. Ausgewähltes Portfolio

Mit den gewichten aus 3.4.1 wurde ein Portfolio aus den Top 20 Aktien ausgewählt (Tabelle1). Es hat sich ergeben, dass das grüne Modul 0 bei weitem den Größten Anteil besitzt.

ID	Name	Cavg	Modul
CFO	Victory CEMP US 500 Enhanced Volatility Wtd Index ETF	152	
VTWO	Vanguard Russell 2000 ETF	129	
VONG	Vanguard Russell 1000 Growth ETF	112	
VTWG	Vanguard Russell 2000 Growth ETF	112	
TBRA	Tobira Therapeutics, Inc.	108	
JKHY	Jack Henry & Associates, Inc.	105	
AAXJ	iShares MSCI All Country Asia ex Japan Index Fund	104	
QCLN	First Trust NASDAQ Clean Edge Green Energy Index Fund	103	
CHW	Calamos Global Dynamic Income Fund	103	
BUSE	First Busey Corporation	103	
ACAD	ACADIA Pharmaceuticals Inc.	102	
CIZ	Victory CEMP Developed Enhanced Volatility Wtd Index ETF	102	
LMCA	Liberty Media Corporation	102	
BANR	Banner Corporation	102	
QVCB	Liberty Interactive Corporation	102	
EFII	Electronics for Imaging, Inc.	102	
JJSF	J & J Snack Foods Corp.	102	
ALNY	Alnylam Pharmaceuticals, Inc.	102	
AVGO	Broadcom Limited	102	
XIV	VelocityShares Daily Inverse VIX Short Term ETN	102	

Tabelle 1: Das ausgewählte Portfolio

Danach wurde die Leistung des Portfolios gemäß dem Vorgehen aus 3.3.4 gemessen (Abbildung 8). Interessant ist zu sehen, dass obwohl das gesamte Modul 0 relativ wenig Übereinstimmung mit der gesamtkurve des Marktes hatte, es jetzt einige der Kursspitzen gut trifft. Insbesondere das Verhalten bei Tag ~110 ist dem darauffolgenden Ausschlag des Gesamtkurses sehr ähnlich und deutet auf eine zeitliche Verschiebung des Einflusses hin. Nichtsdestotrotz wirkt es vielerorts willkürlich und ein deutlicher Vorteil scheint sich nicht herauszukristallisieren.

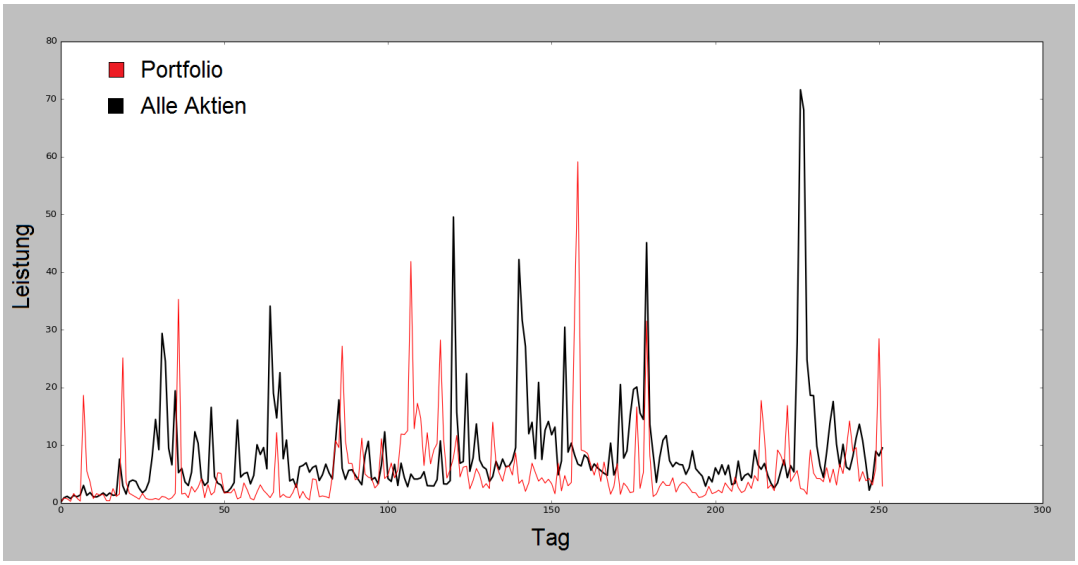

Abbildung 8: Leistung des ausgewählten Portfolios verglichen mit der Leistung aller Aktien

4. Projektabschluss und Aussicht

   Aus heruntergeladenen Börsendaten wurde ein Netzwerk nach dem beschriebenen Verfahren des Papers generiert und ausgewertet. Die Erkenntnisse decken sich mit dem des Papers und es können die gleichen Schlussfolgerungen gezogen werden. Eine ist, dass das Netzwerk bzw. die Module eine gute Repräsentanz der echten Daten darstellen, die andere ist, dass die Formel (1) zur Bestimmung einer Korrelationsmatrix eine Zeitversetzung berücksichtigen sollte, da Kurseinflüsse erst zeitversetzt wirken können.

   Das Optimierungsproblem wurde nicht gelöst, da es über den Projekt- bzw. Kursrahmen gehen würde, weshalb das empirisch ausgewählte Portfolio nicht optimal ist und demnach keinen deutlichen Vorteil erkennen lässt. Dennoch lässt die Auswahl in Abbildung 8 erkennen, wie ein optimales Portfolio mit Zeitverschiebung einen Marktvorteil generieren könnte. Eine marktdefinierende Kurssteigerung, initiiert durch eine kleine Auswahl von Aktien, ist der Optimalfall für einen erfolgreichen Handel. Wenn das Portfolio dabei auch noch auf verschiedene Module aufgeteilt wird, dann ist zudem auch das Risiko minimiert. Das wurde im Paper allerdings nicht genauer ausgeführt. Insgesamt ist der Kern des Projekts erfolgreich durchgeführt, insofern ist das Projekt erfolgreich abgeschlossen.

   Weiterführend könnten die Module zur Vorhersage der Kursänderungen genutzt werden. Tools zur Vorhersage von Kursänderungen werten primär den Kurs einer Aktie aus. Wenn also die Gesamtprognose innerhalb eines Moduls einen überdurchschnittlichen Wert erreicht, so könnte es ein gutes Indiz für eine Handelsaktion sein bzw. die Vorhersagegenauigkeit erhöhen.

   Zur Bewertung des Papers sei gesagt, dass es für mich inhaltlich etwas oberflächlich schien. Der Netzwerkaspekt des Papers sowie dessen Quellen im Bereich der Aktien scheint sich schwerpunktmäßig auf die Konstruktion des Netzwerks zu konzentrieren. Bei den Methoden wurden einige Definitionen gegeben, allerdings wurde bei den Ergebnissen nicht genauer auf die Parameterwahl oder Prozedurbeschreibung eingegangen. Insbesondere sind die Bedeutung und die Argumentation der Zentralitäten, mit einem Satz, deutlich zu kurz. Es war sehr viel Quellenrecherche und Eigeninterpretation erforderlich. Insbesondere der Modulaspekt bzw. der Sinn der *community detection* wurde meiner Meinung nach nicht weit genug ausgeführt.

*Die Formeln (1), (2) sowie aus den Bereichen 3.2.3, 3.3.4 und 3.4.1 sind dem Paper entnommen.