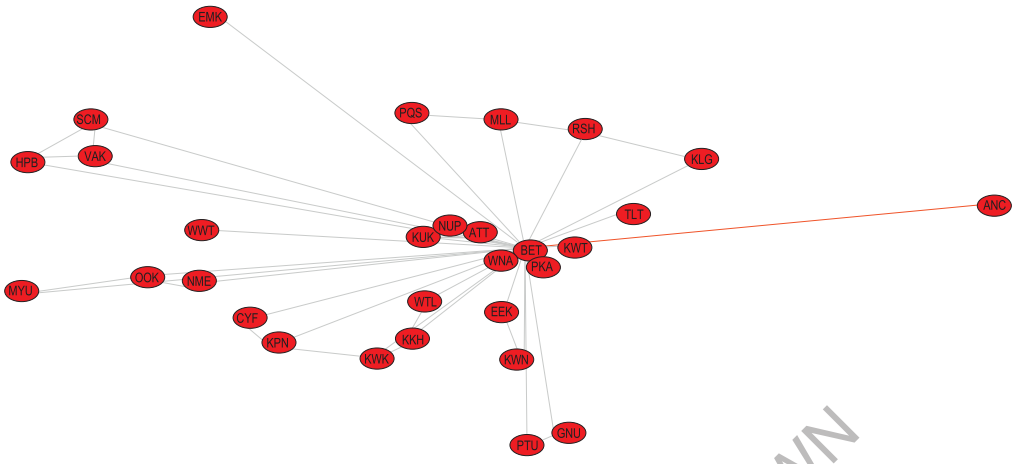


Rysunek 3.9. Sieć ego stopnia 1 połączeń z Bethel w układzie zbliżonym do geograficznego (dla zwiększenia czytelności zmniejszono odległość do Anchorage). Kod IATA Bethel to BET, ANC to Anchorage. Kolor szary oznacza połączenia obsługiwane przez niewielkie (do około 20 miejsc) samoloty, czerwony – samoloty średniej wielkości, np. Boeing 737

się stale między Alaską i Jakucją, to ich trasy w naturalny sposób się krzyżują. Taki stan rzeczy sprzyja powstaniu stosunkowo krótkich połączeń. Inaczej będzie w przypadku miejscowości o mniejszym znaczeniu ekonomicznym. I tak średnica naszego grafu wynosi nie 6 a 13, co ma miejsce na trasie między niewielkim osiedlem na Grenlandii – Aappilattoq (YPO) – a równie niewielką osadą nad kanadyjską zatoką Hudsona – Peawanuck (QUV).

Estymacja parametrów sieci

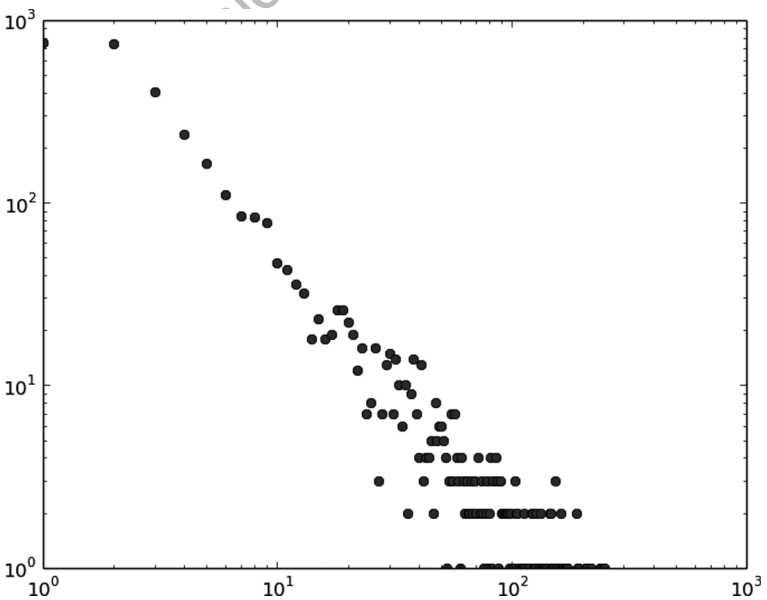
Nawiązując do rozważań dotyczących sieci wybranych przewoźników, należy zadać sobie pytanie, czy tak zbudowaną sieć można uznać za złożoną. W istocie rozmiar sieci zdaje się usprawiedliwiać takie przypuszczenie, jednakże decydujące znaczenie ma nie wielkość, lecz występowanie określonych własności. Najbardziej podstawową jest rozkład wierzchołków. W przypadku sieci Germanwings można było wskazać nieliczne węzły o nietypowo wysokim stopniu. Dla sieci tak dużych rozmiarów jak omawiana, bardziej praktyczne jest sprawdzenie, czy stopnie wierzchołków zmieniają się w sposób dający się opisać jakimś znanym rozkładem. Ponieważ dla sieci złożonej oczekiwanym wynikiem jest rozkład potęgowy, zaczniemy od wykreślenia częstości występowania poszczególnych stopni w skali podwójnie logarytmicznej[29]. Posłużymy się w tym

[29] Wykres taki jest pewnym przybliżeniem funkcji gęstości rozkładu, pozbawionym jednak jakiegokolwiek uśredniania, które mogłoby go wygładzić. Używamy go tutaj, aby uwypuklić dyskretną naturę rozkładu i zademonstrować niebezpieczeństwa związane z takim podejściem. Lepszą metodą jest wykreślenie histogramu z odpowiednio dobranymi przedziałami – najlepiej, aby ich szerokość wykreślona w skali logarytmicznej była jednakowa. Umożliwia to np. metoda plot_pdf z pakietu Pythona powerlaw.

celu kodem zaprezentowanym na wydruku 3.6. Wynik działania programu pokazuje rys. 3.10. Jak widać, dla węzłów niskiego stopnia (czyli niewielkich lotnisk) punkty układają się w przybliżeniu wzdłuż linii prostej, jednakże punkty odpowiadające wielkim lotniskom są dość mocno rozsiane. Wydaje się to dość oczywiste, gdyż ich stosunkowo niewielka liczba wyklucza uśrednianie. Tak czy inaczej, rozkład punktów utrudnia weryfikację hipotezy o potęgowości rozkładu. Bez przeprowadzania obliczeń łatwo zauważyć, że ze względu na rozszerzający się kształt wypełnionego punktami obszaru, można próbować dopasować do nich proste o różnym nachyleniu (odpowiadające rozkładom potęgowym o różnym wykładniku α), a nawet linie krzywe.

```
1 from collections import Counter
2 deg_dict = Counter(dlist)
3 degs = deg_dict.keys();
4 freqs = deg_dict.values()
5 plt.loglog(degs, freqs, 'bo')
6 plt.gcf().savefig('degree_freq.png')
7 plt.close()
```

Wydruk 3.6. Wykreślenie częstości występowania stopni węzłów

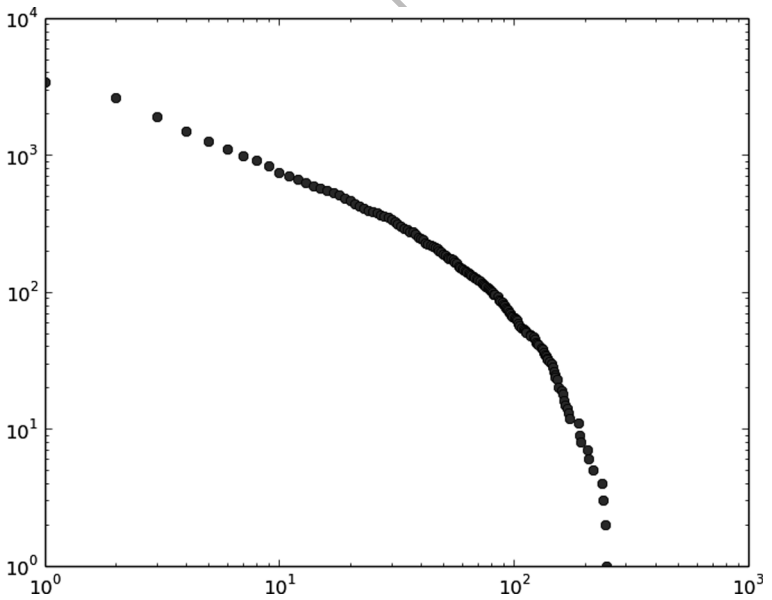


Rysunek 3.10. Wykres częstości występowania stopni węzłów

Lepszym rozwiązaniem jest wykreślenie wykresu rankingowego (patrz podrozdz. 3.1). Kod programu[30] przedstawia wydruk 3.7. Wynik działania programu widać na rys. 3.11. Wykres ten rzuca nieco światła na naturę obserwowanego rozkładu. Nie można stwierdzić, aby w całym zakresie zmienności stopnia węzłów był on rozkładem potęgowym. Widać wyraźnie, że mimo użycia wykresu rankingowego nachylenie nadal zmienia się w końcowym odcinku rozkładu. Należy więc odpowiedzieć sobie na pytanie: czy większe znaczenie ma dla nas ten końcowy odcinek, czy też rozkład stopnia wszystkich węzłów? W pierwszym przypadku możemy ograniczyć się do np. ostatniej dekady wykresu i próbować dopasować do niej rozkład potęgowy o znacznie większym wykładniku α niż dla danych mieszczących się w początkowym fragmencie.

```
cumul=[]
for i in range(len(degs)):
    cnt = 0
    for j in range(i,len(degs)):
        cnt = cnt + freqs[j]
    cumul.append(cnt)
plt.loglog(degs, cumul, 'bo')
```

Wydruk 3.7. Wyznaczenie wykresu rangowego stopni węzłów



Rysunek 3.11. Wykres rangowy stopni węzłów

[30] Podobnie jak w poprzednim przypadku dla przejrzystości i uwypuklenia dyskretnej natury rozkładu zdecydowaliśmy się na własną implementację. W praktyce z powodzeniem można wykorzystać metodę plot_ccdf pakietu powerlaw.

W celu estymacji współczynnika α najlepiej posłużyć się metodą największej wiarygodności [38]. W przypadku ciągłego rozkładu potęgowego możliwe jest znalezienie analitycznego rozwiązania w postaci:

$$\hat{\alpha} = 1 + \frac{n}{\sum_{i=0}^{n} \log\left(\frac{x_i}{x_{\min}}\right)}, \tag{3.3}$$

gdzie $x_{\min}$ oznacza minimalną wartość zmiennej losowej. Stosowanie tego estymatora w przypadku dyskretnym powoduje powstanie błędu, tym większego, im mniejsze jest $x_{\min}$. Z tego powodu w przypadku dyskretnym lepiej stosować bardziej czasochłonną[31] procedurę iteracyjną [38], implementowaną między innymi przez metodę `fit` pakietu powerlaw [7]. Aby dobrać właściwą wartość $x_{\min}$, wskazane byłoby przeprowadzenie szeregu eksperymentów pozwalających ocenić, czy wartość $\hat{\alpha}$ stabilizuje się z wzrostem $x_{\min}$. Jeżeli tak się dzieje, to można przypuszczać, że rozkład potęgowy jest właściwym wyborem, w przeciwnym przypadku można podejrzewać, że dane mają inny rozkład. Korzystając z metody `fit` z opcją `discrete = True`, w stosunkowo prosty sposób można wyznaczyć estymaty współczynnika rozkładu potęgowego dla całego zakresu zmienności stopnia węzłów, otrzymując tzw. *wykres Hilla* – patrz wydruk 3.8 i rys. 3.12. Niestety wykres ten nie wykazuje oczekiwanego zachowania – zamiast ustalać się w końcowej części, krzywa gwałtownie rośnie. W istocie rzeczy zjawisko to można dość łatwo wytłumaczyć, nie wnikając w szczegóły rozkładu stopnia wierzchołków, a kierując się jedynie wcześniejszymi obserwacjami. Zauważyliśmy bowiem, że decydujące znaczenie dla sposobu, w jaki działa komunikacja lotnicza, mają wielkie lotniska, tzw. huby, w których koncentrują się loty z podobnych lotnisk całego świata, ale również krótsze połączenia regionalne. Lotnisk tych jest jednak ograniczona liczba, co kłóci się z wykorzystaniem rozkładu potęgowego, który z natury swojej jest zdefiniowany na przedziale nieskończonym.

Ta właściwość grafu może wskazywać, że lepszym kandydatem byłby obcięty rozkład potęgowy, czyli rozkład opisany następującą funkcją gęstości:

$$P(x) = \frac{\alpha - 1}{x_{\min}^{1-\alpha} - x_{\max}^{1-\alpha}} x^{-\alpha}, \tag{3.4}$$

gdzie $x_{\min}$ oznacza minimalną a $x_{\max}$ maksymalną wartość zmiennej losowej. Wzór (3.4) jest prawdziwy dla $x_{\min} > 0$. Warto zauważyć, że gdy $x_{\max} \to \infty$, wzór (3.4) staje się rozkładem potęgowym, czyli p_0 z równania (3.1) jest równe $\frac{\alpha - 1}{x_{\min}^{1-\alpha}}$. Aby sprawdzić hipotezę o takiej postaci rozkładu stopni węzłów naszego grafu, dopasujemy do danych rozkład potęgowy oraz rozkłady: obcięty potęgowy i lognormalny – kod tej operacji przedstawia wydruk 3.9. Użycie rozkładu lognormalnego jest wynikiem doświadczeń: często jest on mylony z rozkładem potęgowym. W istocie jest on również rozkładem ciężkoogonowym. Wynikowy wykres przestawia rys. 3.13. Jak widać, dopasowanie rozkładu

[31] Zgodnie z [38] wartość wykładnika α można w przypadku dyskretnym przybliżyć, podstawiając do równania (3.3) $x_{\min} - \frac{1}{2}$ w miejsce $x_{\min}$.

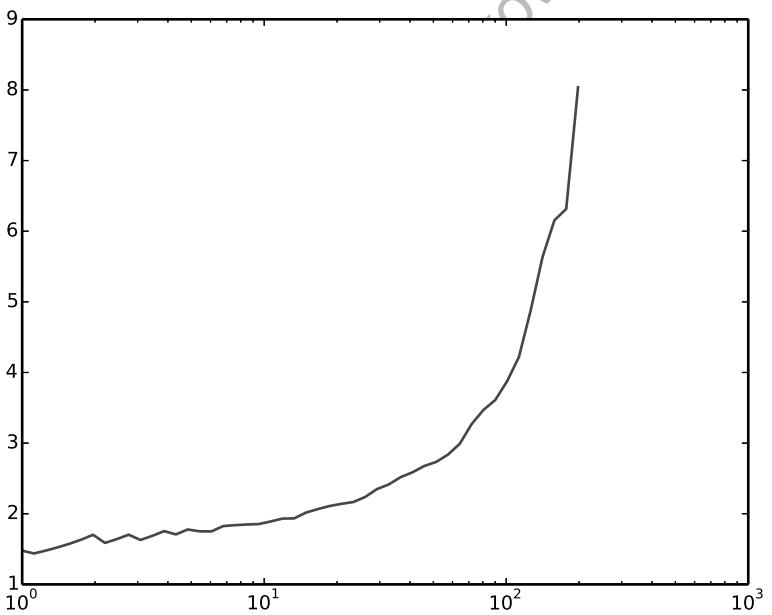
```
import powerlaw
import numpy as np

NBINS=50
bins=np.logspace(np.log10(min(dlist)),
                 np.log10(max(dlist)), num=NBINS)
bcnt,bedge=np.histogram(np.array(dlist),bins=bins)
alpha=np.zeros(len(bedge[:-2]))

for i in range(0,len(bedge)-2):
    fit=powerlaw.Fit(dlist,xmin=bedge[i], discrete=True)
    alpha[i]=fit.alpha

plt.semilogx(bedge[:-2],alpha)  # Hill plot
```

Wydruk 3.8. Wyznaczenie wykresu Hilla



Rysunek 3.12. Wykres Hilla stopni węzłów

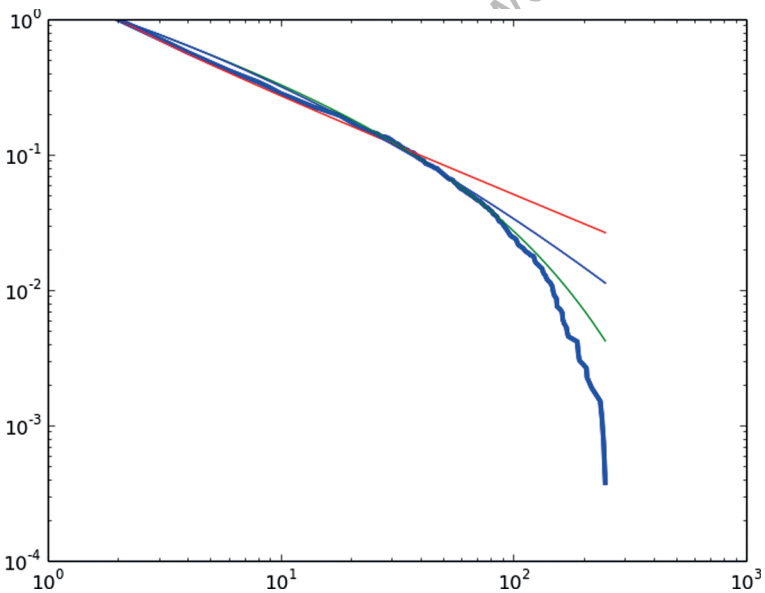
potęgowego w całym zakresie zmienności stopnia węzłów[32] skutkuje uwzględnieniem głównie węzłów o niskim stopniu, z czego wynika stosunkowo niewielka wartość wykładnika α ($\alpha = 1{,}71$). Linia o tak małym nachyleniu jest bardzo odległa od danych

[32] Jest to wynik działania funkcji `fit` bez dodatkowego parametru `xmin`.

dla wysokich stopni węzłów. Zastosowanie rozkładu lognormalnego pozwala nieco zmniejszyć ten błąd, jednakże znacznie lepszy (choć nadal nie idealny) jest obcięty rozkład potęgowy.

```
fit=powerlaw.Fit(dlist, discrete=True)
print 'alpha (disc.) = ' + str(fit.alpha)
fig3 = fit.plot_ccdf(linewidth = 3)
fit.power_law.plot_ccdf(ax = fig3, color = 'r',
                        linestyle = '-')
fit.truncated_power_law.plot_ccdf(ax = fig3, color = 'g',
                                  linestyle = '-')
fit.lognormal.plot_ccdf(ax = fig3, color = 'b',
                        linestyle = '-')
```

Wydruk 3.9. Porównanie rozkładów potęgowego z obciętym potęgowym i lognormalnym



Rysunek 3.13. Rozkład potęgowy (linia czerwona), obcięty potęgowy (linia zielona) i lognormalny (linia niebieska) na tle danych (pogrubiona linia niebieska).

Aby upewnić się o trafnym doborze rozkładu, można sprawdzić hipotezę o jego lepszym dopasowaniu niż innego wskazanego rozkładu. Wydruk 3.10 pokazuje, jak sprawdzić dopasowanie rozkładu potęgowego w porównaniu z rozkładami: wykładniczym, obciętym potęgowym i lognormalnym. Otrzymane wyniki zawiera tabela 3.2. We wszystkich przypadkach małe p-wartości świadczą o małej wariancji błędów i co za tym idzie – wiarygodności wyników testu. Pierwszy test – porównanie z rozkładem

wykładniczym – pokazuje, że ten rozkład jest zdecydowanie gorzej dopasowany. Wskazuje to na ciężkoogonowość, której rozkład ten nie może oddać. W odróżnieniu od pierwszego dwa pozostałe testy sugerują zmianę rozkładu, przy czym, analogicznie jak na rys. 3.13, obcięty rozkład potęgowy wypada najlepiej. Może to świadczyć o tym, że sieć połączeń lotniczych jest w gruncie rzeczy typową siecią złożoną, a obserwowane nieregularności rozkładu są tylko wynikiem jej ograniczonej wielkości.

```
R, p = fit.distribution_compare('power_law', 'exponential',
                                normalized_ratio = True)
print 'power_law ? exponential: ' + str(R) + " " + str(p)

R, p = fit.distribution_compare('power_law',
                                'truncated_power_law',
                                normalized_ratio = True)
print 'power_law ? truncated_power_law: ' +
                        str(R) + " " + str(p)

R, p = fit.distribution_compare('power_law', 'lognormal',
                                normalized_ratio = True)

print 'power_law ? lognormal: ' + str(R) + " " + str(p)
```

Wydruk 3.10. Porównanie dopasowania rozkładu potęgowego z rozkładami: wykładniczym, obciętym potęgowym i lognormalnym

Tabela 3.2. Porównanie dopasowanego rozkładu potęgowego z przykładowymi rozkładami

Rozkład	Statystyka R	p-wartość
wykładniczy	17,42	$6{,}2 \cdot 10^{-68}$
obcięty potęgowy	−10,94	0
lognormalny	−7,34	$2{,}1 \cdot 10^{-13}$

3.4. Wielkości opisujące sieć

Oprócz fundamentalnych pojęć bezskalowości oraz małych światów w celu opisu sieci złożonych definiuje się również inne wskaźniki, przedstawione poniżej. Uzupełniają one zestaw definicji podanych w podrozdz. 2.2 i dobrze wpisują się w prezentację sieci jako struktur obiektów traktowanych podmiotowo – choć oczywiście mogą być i są stosowane w opisie „zwykłych” grafów.