CSYS 300 Assignment 5

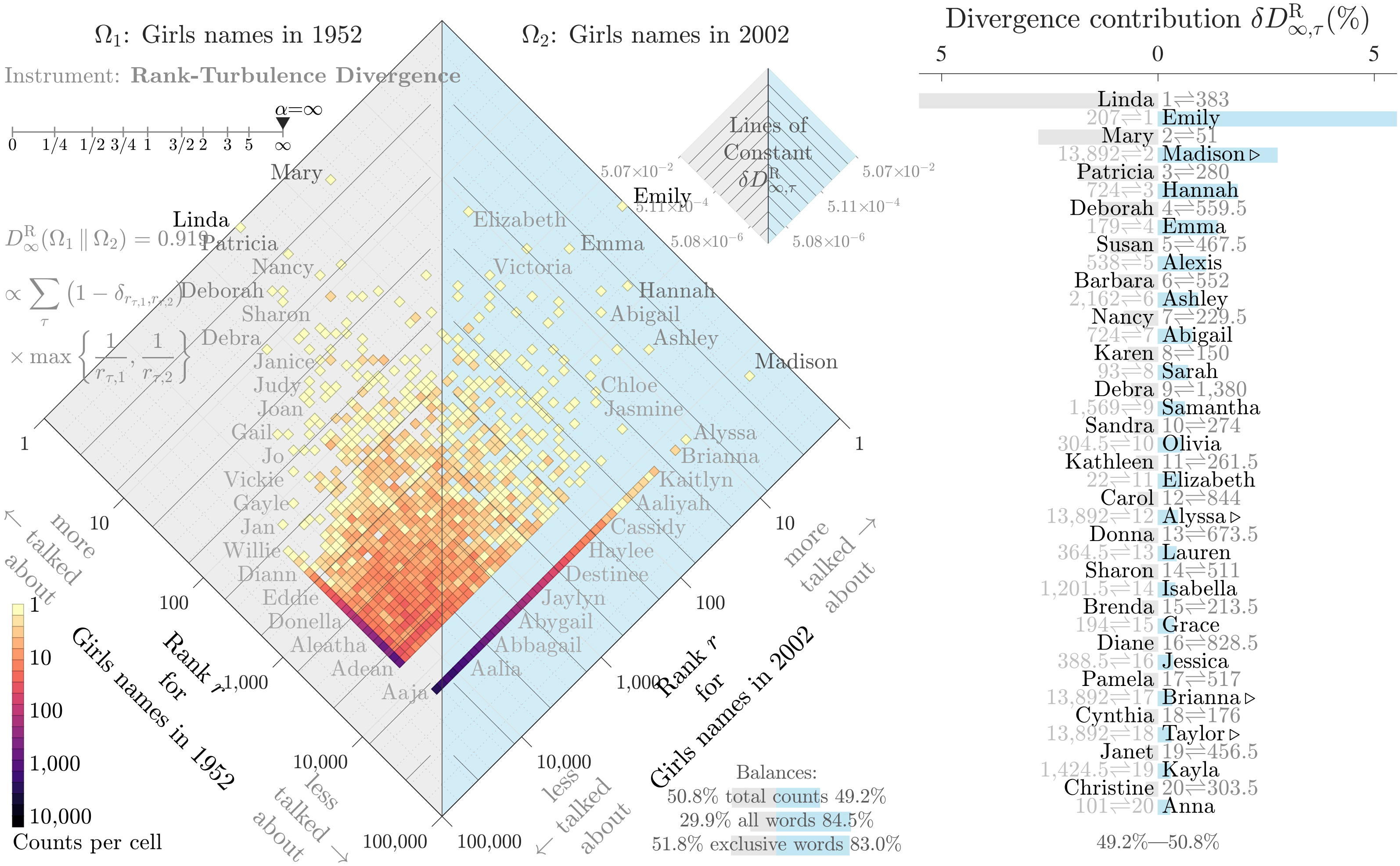
Missy Westland (mwestlan@uvm.edu)

October 1, 2021

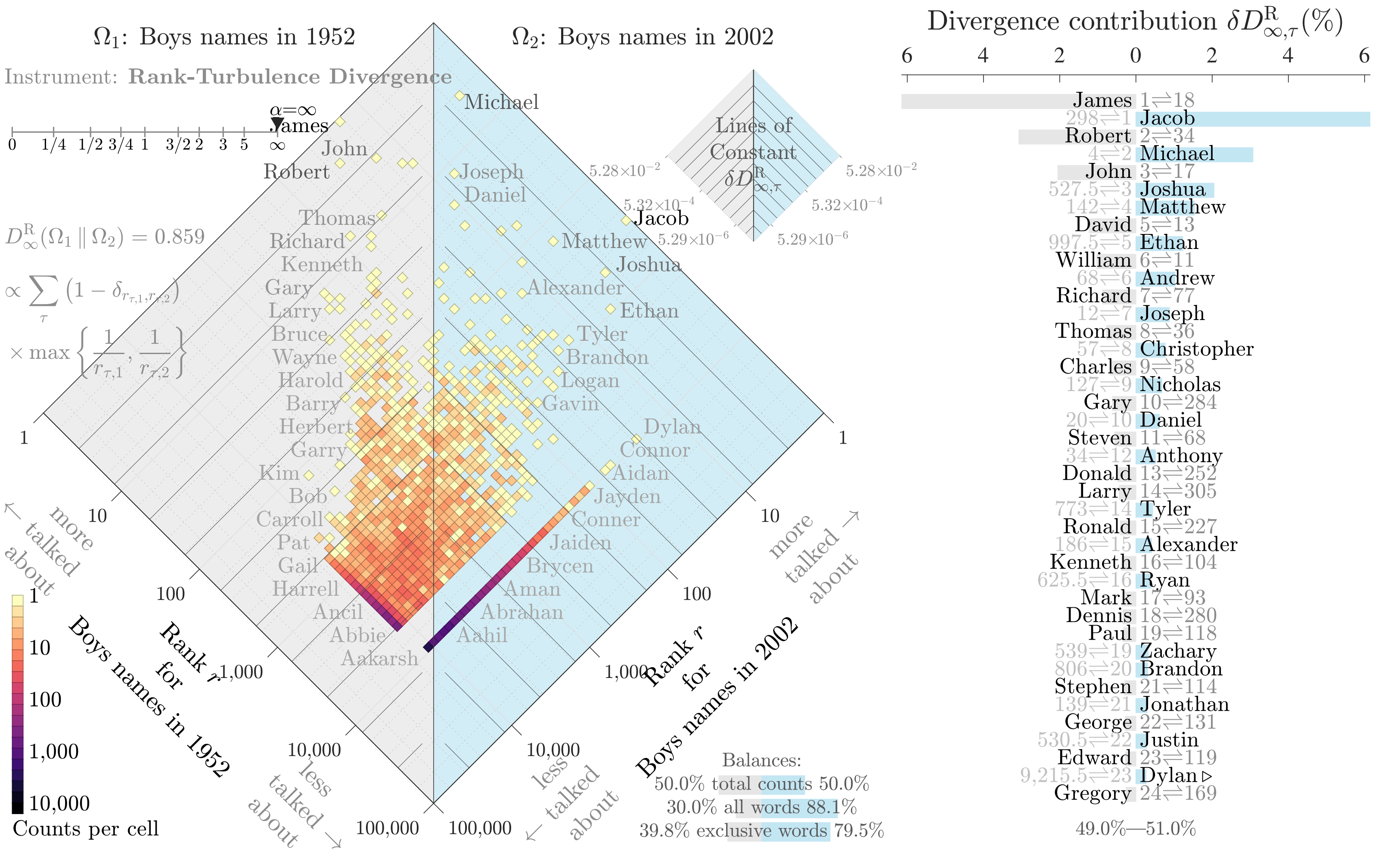
Link to Github for Assignment Code: `YOURLINKTOYOURGITHUB`

Question 1.

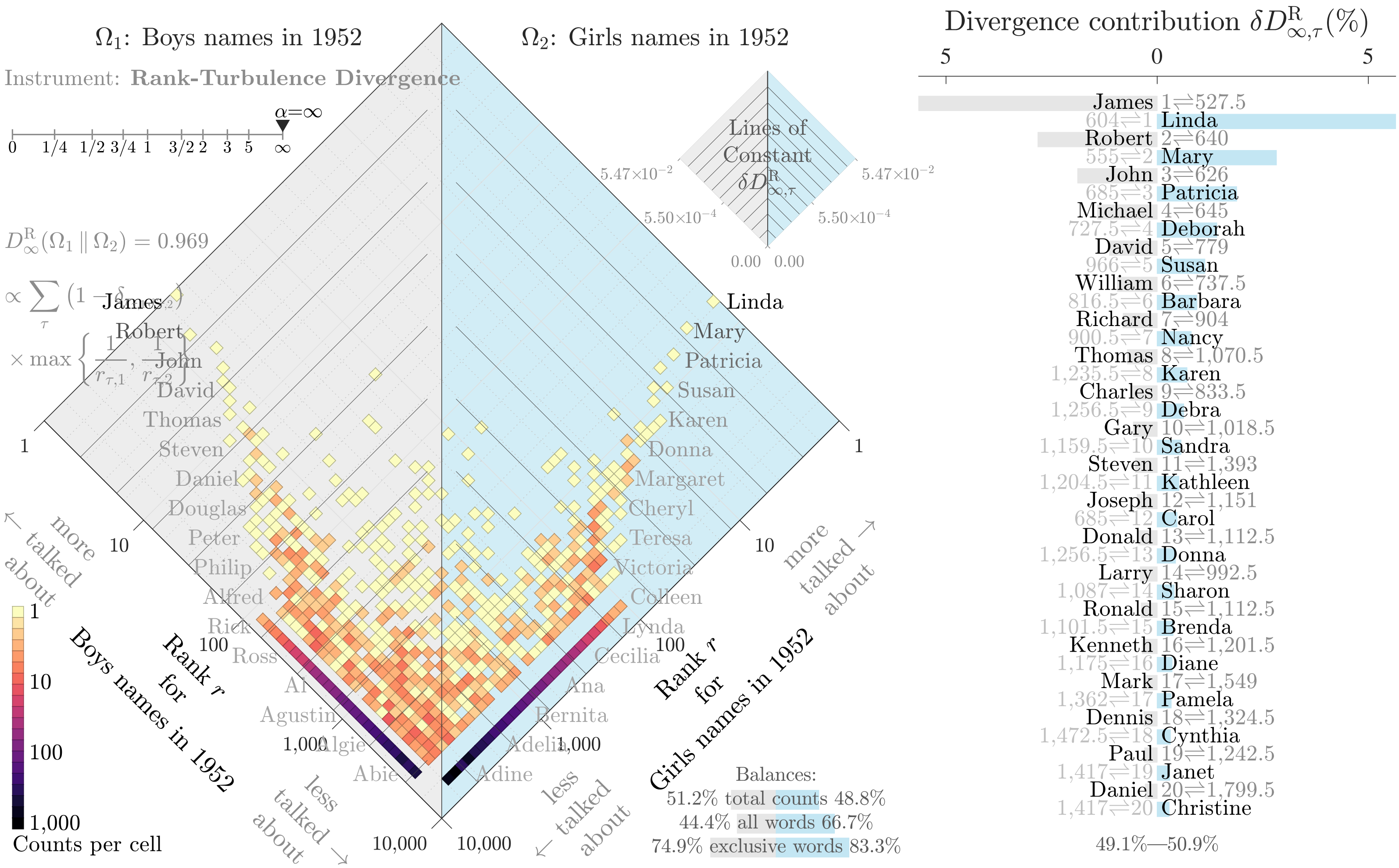
Generate allotaxonographs comparing the following four pairs: **Responses:**



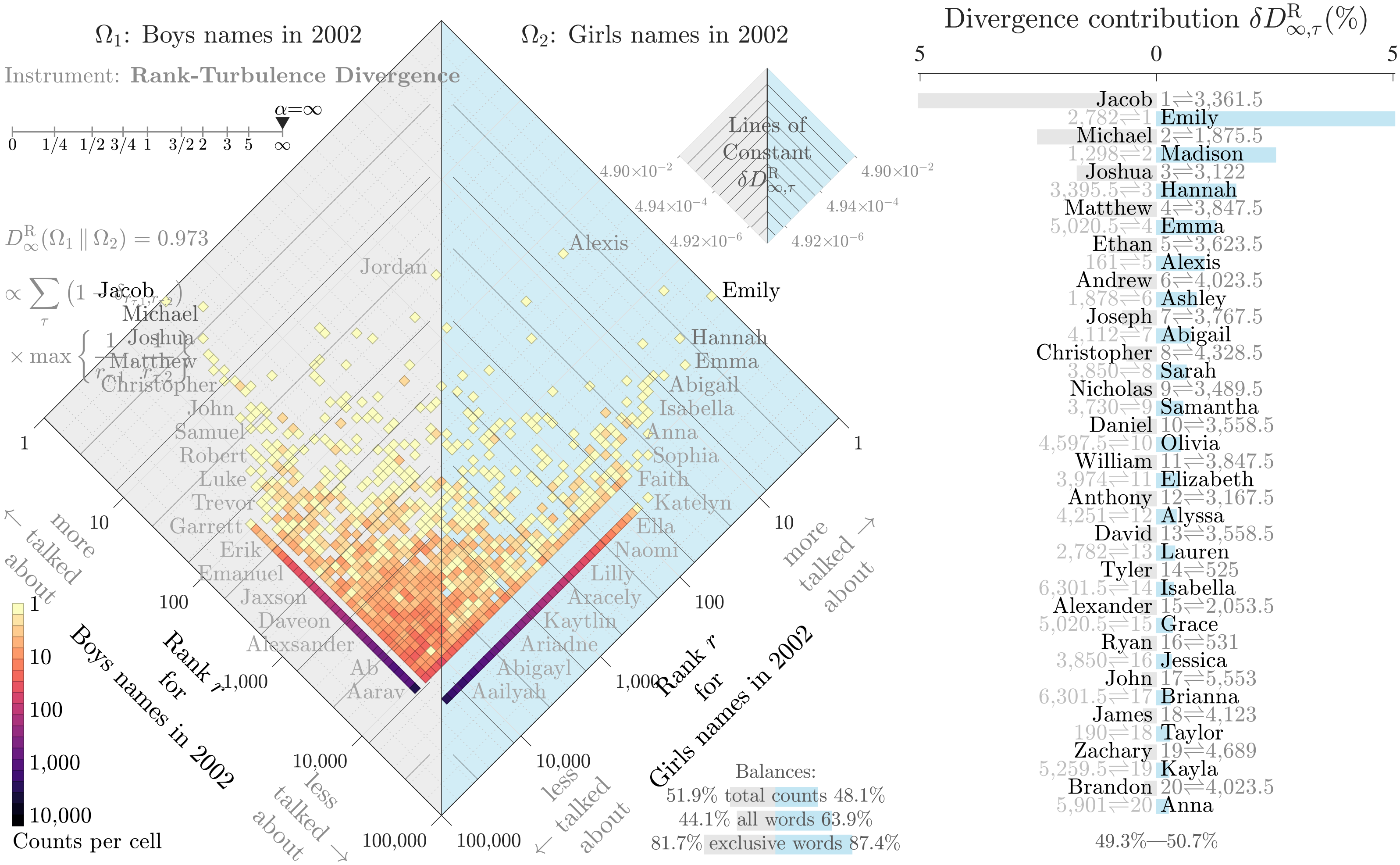
Ω_1: Girls names in 1952
Ω_2: Girls names in 2002
Instrument: **Rank-Turbulence Divergence**
$\alpha=\infty$
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D^{\mathrm{R}}_{\infty}(\Omega_1 \,\|\, \Omega_2) = 0.919$
$\propto \sum_{\tau} \left(1 - \delta_{r_{\tau,1}, r_{\tau,2}}\right) \times \max\left\{\frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}}\right\}$
Lines of Constant $\delta D^{\mathrm{R}}_{\infty,\tau}$
5.07×10^{-2}
5.11×10^{-4}
5.08×10^{-6}
Mary
Linda
Patricia
Nancy
Deborah
Sharon
Debra
Janice
Judy
Joan
Gail
Jo
Vickie
Gayle
Jan
Willie
Diann
Eddie
Donella
Aleatha
Adean
Aaja
Emily
Elizabeth
Emma
Victoria
Hannah
Abigail
Ashley
Madison
Chloe
Jasmine
Alyssa
Brianna
Kaitlyn
Aaliyah
Cassidy
Haylee
Destinee
Jaylyn
Abygail
Abbagail
Aalia
1
10
100
1,000
10,000
100,000
← more talked about
less talked about →
more talked about →
← less talked about
Rank r for Girls names in 1952
Rank r for Girls names in 2002
1
10
100
1,000
10,000
Counts per cell
Balances:
50.8% total counts 49.2%
29.9% all words 84.5%
51.8% exclusive words 83.0%
Divergence contribution $\delta D^{\mathrm{R}}_{\infty,\tau}(\%)$
5 0 5
Linda 1⇌383
207⇌1 Emily
Mary 2⇌51
13,892⇌2 Madison ▷
Patricia 3⇌280
724⇌3 Hannah
Deborah 4⇌559.5
179⇌4 Emma
Susan 5⇌467.5
538⇌5 Alexis
Barbara 6⇌552
2,162⇌6 Ashley
Nancy 7⇌229.5
724⇌7 Abigail
Karen 8⇌150
93⇌8 Sarah
Debra 9⇌1,380
1,569⇌9 Samantha
Sandra 10⇌274
304.5⇌10 Olivia
Kathleen 11⇌261.5
22⇌11 Elizabeth
Carol 12⇌844
13,892⇌12 Alyssa ▷
Donna 13⇌673.5
364.5⇌13 Lauren
Sharon 14⇌511
1,201.5⇌14 Isabella
Brenda 15⇌213.5
194⇌15 Grace
Diane 16⇌828.5
388.5⇌16 Jessica
Pamela 17⇌517
13,892⇌17 Brianna ▷
Cynthia 18⇌176
13,892⇌18 Taylor ▷
Janet 19⇌456.5
1,424.5⇌19 Kayla
Christine 20⇌303.5
101⇌20 Anna
49.2%—50.8%



Ω1: Boys names in 1952
Ω2: Boys names in 2002
Instrument: Rank-Turbulence Divergence
α=∞
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
D∞R(Ω1 ‖ Ω2) = 0.859
∝ Σ_τ (1 − δ_{r_τ,1, r_τ,2}) × max{1/r_τ,1, 1/r_τ,2}
Lines of Constant δD∞,τR
5.28×10−2
5.32×10−4
5.29×10−6
James
John
Robert
Thomas
Richard
Kenneth
Gary
Larry
Bruce
Wayne
Harold
Barry
Herbert
Garry
Kim
Bob
Carroll
Pat
Gail
Harrell
Ancil
Abbie
Aakarsh
Michael
Joseph
Daniel
Jacob
Matthew
Joshua
Alexander
Ethan
Tyler
Brandon
Logan
Gavin
Dylan
Connor
Aidan
Jayden
Conner
Jaiden
Brycen
Aman
Abrahan
Aahil
1
10
100
1,000
10,000
100,000
more talked about
less talked about
Rank r for Boys names in 1952
Rank r for Boys names in 2002
Counts per cell
1
10
100
1,000
10,000
Balances:
50.0% total counts 50.0%
30.0% all words 88.1%
39.8% exclusive words 79.5%
Divergence contribution δD∞,τR (%)
6 4 2 0 2 4 6
James 1⇌18
298⇌1 Jacob
Robert 2⇌34
4⇌2 Michael
John 3⇌17
527.5⇌3 Joshua
142⇌4 Matthew
David 5⇌13
997.5⇌5 Ethan
William 6⇌11
68⇌6 Andrew
Richard 7⇌77
12⇌7 Joseph
Thomas 8⇌36
57⇌8 Christopher
Charles 9⇌58
127⇌9 Nicholas
Gary 10⇌284
20⇌10 Daniel
Steven 11⇌68
34⇌12 Anthony
Donald 13⇌252
Larry 14⇌305
773⇌14 Tyler
Ronald 15⇌227
186⇌15 Alexander
Kenneth 16⇌104
625.5⇌16 Ryan
Mark 17⇌93
Dennis 18⇌280
Paul 19⇌118
539⇌19 Zachary
806⇌20 Brandon
Stephen 21⇌114
139⇌21 Jonathan
George 22⇌131
530.5⇌22 Justin
Edward 23⇌119
9,215.5⇌23 Dylan ▷
Gregory 24⇌169
49.0%—51.0%



Ω1: Boys names in 1952
Ω2: Girls names in 1952
Instrument: Rank-Turbulence Divergence
α=∞
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D^{\mathrm{R}}_{\infty}(\Omega_1 \,\|\, \Omega_2) = 0.969$
$\propto \sum_\tau (1 - \delta_{\tau,\Omega_1\Omega_2}) \times \max\left\{\frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}}\right\}$
Lines of Constant $\delta D^{\mathrm{R}}_{\infty,\tau}$
5.47×10^-2 5.50×10^-4 0.00
James
Robert
John
David
Thomas
Steven
Daniel
Douglas
Peter
Philip
Alfred
Rick
Ross
Al
Agustin
Algie
Abie
Linda
Mary
Patricia
Susan
Karen
Donna
Margaret
Cheryl
Teresa
Victoria
Colleen
Lynda
Cecilia
Ana
Bernita
Adelia
Adine
1 10 100 1,000 10,000
← more talked about
less talked about →
more talked about →
← less talked about
Rank r for Boys names in 1952
Rank r for Girls names in 1952
1 10 100 1,000
Counts per cell
Balances:
51.2% total counts 48.8%
44.4% all words 66.7%
74.9% exclusive words 83.3%
Divergence contribution $\delta D^{\mathrm{R}}_{\infty,\tau}$(%)
5 0 5
James 1⇌527.5
604⇌1 Linda
Robert 2⇌640
555⇌2 Mary
John 3⇌626
685⇌3 Patricia
Michael 4⇌645
727.5⇌4 Deborah
David 5⇌779
966⇌5 Susan
William 6⇌737.5
816.5⇌6 Barbara
Richard 7⇌904
900.5⇌7 Nancy
Thomas 8⇌1,070.5
1,235.5⇌8 Karen
Charles 9⇌833.5
1,256.5⇌9 Debra
Gary 10⇌1,018.5
1,159.5⇌10 Sandra
Steven 11⇌1,393
1,204.5⇌11 Kathleen
Joseph 12⇌1,151
685⇌12 Carol
Donald 13⇌1,112.5
1,256.5⇌13 Donna
Larry 14⇌992.5
1,087⇌14 Sharon
Ronald 15⇌1,112.5
1,101.5⇌15 Brenda
Kenneth 16⇌1,201.5
1,175⇌16 Diane
Mark 17⇌1,549
1,362⇌17 Pamela
Dennis 18⇌1,324.5
1,472.5⇌18 Cynthia
Paul 19⇌1,242.5
1,417⇌19 Janet
Daniel 20⇌1,799.5
1,417⇌20 Christine
49.1%—50.9%



Ω_1: Boys names in 2002
Ω_2: Girls names in 2002
Instrument: **Rank-Turbulence Divergence**
α=∞
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D^{\mathrm{R}}_{\infty}(\Omega_1 \,\|\, \Omega_2) = 0.973$
Lines of Constant $\delta D^{\mathrm{R}}_{\infty,\tau}$
4.90×10^{-2}
4.94×10^{-4}
4.92×10^{-6}
Jacob
Michael
Joshua
Matthew
Christopher
John
Samuel
Robert
Luke
Trevor
Garrett
Erik
Emanuel
Jaxson
Daveon
Alexsander
Ab
Aarav
Jordan
Alexis
Emily
Hannah
Emma
Abigail
Isabella
Anna
Sophia
Faith
Katelyn
Ella
Naomi
Lilly
Aracely
Kaytlin
Ariadne
Abigayl
Aailyah
1
10
100
1,000
10,000
100,000
← more talked about
more talked about →
less talked about →
← less talked about
Rank r for Boys names in 2002
Rank r for Girls names in 2002
1
10
100
1,000
10,000
Counts per cell
Balances:
51.9% total counts 48.1%
44.1% all words 63.9%
81.7% exclusive words 87.4%
Divergence contribution $\delta D^{\mathrm{R}}_{\infty,\tau}$(%)
5 0 5
Jacob 1⇄3,361.5
2,782⇄1 Emily
Michael 2⇄1,875.5
1,298⇄2 Madison
Joshua 3⇄3,122
3,395.5⇄3 Hannah
Matthew 4⇄3,847.5
5,020.5⇄4 Emma
Ethan 5⇄3,623.5
161⇄5 Alexis
Andrew 6⇄4,023.5
1,878⇄6 Ashley
Joseph 7⇄3,767.5
4,112⇄7 Abigail
Christopher 8⇄4,328.5
3,850⇄8 Sarah
Nicholas 9⇄3,489.5
3,730⇄9 Samantha
Daniel 10⇄3,558.5
4,597.5⇄10 Olivia
William 11⇄3,847.5
3,974⇄11 Elizabeth
Anthony 12⇄3,167.5
4,251⇄12 Alyssa
David 13⇄3,558.5
2,782⇄13 Lauren
Tyler 14⇄525
6,301.5⇄14 Isabella
Alexander 15⇄2,053.5
5,020.5⇄15 Grace
Ryan 16⇄531
3,850⇄16 Jessica
John 17⇄5,553
6,301.5⇄17 Brianna
James 18⇄4,123
190⇄18 Taylor
Zachary 19⇄4,689
5,259.5⇄19 Kayla
Brandon 20⇄4,023.5
5,901⇄20 Anna
49.3%—50.7%

Question 2.

Everyday random walks and the Central Limit Theorem:
Show that the observation that the number of discrete random walks of duration $t = 2n$ starting at $x_0 = 0$ and ending at displacement $x_2 n = 2k$ where $k \epsilon (0, \pm 1, \pm 2, ..., \pm n)$ is

$$N(0, 2k, 2n) = (\begin{matrix} 2n \\ n+k \end{matrix}) = (\begin{matrix} 2n \\ n-k \end{matrix})$$

leads to a Gaussian distribution for large $t = 2n$:

$$\mathbf{Pr}(x_t \equiv x) \simeq \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}$$

Please note that $k \ll n$.
Stirling's sterling approximation will prove most helpful.
Hint: You should be able to reach this form:

$$\frac{\text{Some stuff not involving penguins}}{\text{Some other penguin-free stuff} \times (1 - k^2/n^2)^{n+1/2} (1 + k/n)^k (1 - k/n)^{-k}}$$

Lots of sneakiness here. You'll want to examine the natural log of the piece shown above, and see how it behaves for large n.
You may very well need to use the Taylor expansion $ln(1 + z) \equiv z$.
Exponentiate and carry on.

Tip: If at any point penguins appear in your expression, you're in real trouble. Get some fresh air and start again.

Responses:

$$\frac{(2n)!}{(n+k)!(n-k)!}$$
$$= \frac{\sqrt{2\pi}}{\sqrt{2\pi}^2} * \frac{(2n)^{2n+1/2}e^{-2n}}{(n+k)^{n+k+1/2}(n-k)^{n-k+1/2}e^{-(n+k)+-(n-k)}}$$
$$= \frac{1}{\sqrt{2\pi}} * \frac{(2n)^{2n+1/2}}{(n+k)^{n+k+1/2}(n-k)^{n-k+1/2}}$$
$$= \frac{1}{\sqrt{2\pi}} * \frac{(2n)^{2n+1/2}}{(n(1+\frac{k}{n}))^{n+k+1/2}(n(1-\frac{k}{n}))^{n-k+1/2}}$$
$$= \frac{1}{\sqrt{2\pi n}} * \frac{(2)^{2n+1/2}}{(1+\frac{k}{n})^{n+k+1/2}(1-\frac{k}{n})^{n-k+1/2}}$$
$$= \frac{1}{\sqrt{2\pi n}} * \frac{(2)^{2n+1/2}}{(1-k^2/n^2)^{n+1/2}(1+\frac{k}{n})^{k}(1-\frac{k}{n})^{-k}}$$
$$= (2n)\ln(2^{1/2}) - \ln(\sqrt{2\pi n}) - (n+1/2)\ln(1-\frac{k^2}{n^2}) + k\ln(1+\frac{k}{n}) - k\ln(1-\frac{k}{n})] \quad \text{then use Taylor expansion}$$
$$= (2n)\ln(2^{1/2}) - \ln(\sqrt{2\pi n}) + (n+1/2)(\frac{k^2}{n^2}) - k\frac{k}{n} - k(\frac{k}{n})$$
$$= (2n)\ln(2^{1/2}) - \ln(\sqrt{2\pi n}) - \frac{k^2}{n} - \frac{k^2}{2n^2} + k\frac{k}{n} - k\frac{k}{n}$$
$$= (2n)\ln(2^{1/2}) - \ln(\sqrt{2\pi n}) - \frac{k^2}{n}$$
$$= \exp 2n\ln(2^{1/2}) - \ln(\sqrt{2\pi n}) - \frac{k^2}{n^2}$$
$$= \frac{2^{1/2}}{\sqrt{2\pi n}}e^{\frac{-k^2}{n}} = \ldots \textbf{I know I'm somewhat close, just running out of time}$$

Question 3.

From the lectures, show that the number of distinct 1-d random walk that start at $x = i$ and end at $x = j$ after t time steps is

$$N(i,j,t) = (\frac{t}{(t+j-i)/2})$$

Assume that j is reachable from i after t time steps.

Responses:

We know that the number of random walks is $N_{distinct} = \binom{t}{N}$
.

$$P + N = t$$
$$P - N = j - i$$
$$\implies P = t - N$$
$$\implies t - N - N = j - i$$
$$N = \frac{t-j+i}{2}$$
$$\implies \binom{t}{N} = \binom{t}{\frac{t-j+i}{2}}$$

Question 4.

Discrete random walks:
In class, we argued that the number of random walks returning to the origin for the first time after 2n time steps is given by $N_{\text{first return}}(2n) = N_{\text{fr}}(2n) = N(1, 1, 2n-2) - N(-1, 1, 2n-2)$ where $N(i, j, t) = \binom{t}{(t+j-i)/2}$
Find the leading order term for $N_{\text{fr}}(2n)$ as $n \to \infty$.
Two-step approach:
(a) Combine the terms to form a single fraction,
(b) and then again use Stirling's bonza approximation

If you enjoy this sort of thing, you may like to explore the same problem for random walks in higher dimensions. Seek out George Pólya.
And we are connecting to much other good stuff in combinatorics; more to come in the solutions.

Responses:

$$\begin{aligned}
\binom{2n-2}{(2n-2+1)/2} - \binom{2n-2}{(2n-2+1+1)/2} &= \binom{2n-2}{n-1} - \binom{2n-2}{n} \\
&= \frac{(2n-2)!}{(n-1)!(n-1)!} - \frac{(2n-2)!}{n!(n-2)!} \\
&= \frac{(2n-2)!}{(n-1)!(n-1)!} - \frac{(2n-2)!(n-1)}{n(n-1)!(n-1)!} \\
&= \frac{n(2n-2)!}{n(n-1)!(n-1)!} - \frac{(2n-2)!(n-1)}{n(n-1)!(n-1)!} \\
&= \frac{(2n-2)!(n-(n-1))}{n(n-1)!(n-1)!} \\
&= \frac{(2n-2)!}{n(n-1)!(n-1)!}
\end{aligned}$$

Now apply Stirling's formula

$$\begin{aligned}
&= \frac{1}{\sqrt{2\pi}} \frac{(2n-2)^{2n-3/2} e^{2-2n}}{n(n-1)^{2(n-1/2)} e^{-2(n-1)}} \\
&= \frac{1}{\sqrt{2\pi}} \frac{(2n-2)^{2n-3/2}}{n(n-1)^{2n-1}} \\
&= \frac{1}{\sqrt{2\pi}} \frac{2^{2n-3/2}}{n(n-1)^{1/2}}
\end{aligned}$$

The leading term in this as $n \to \infty$ is $2^{2n-3/2}$

Question 5.

More on the peculiar nature of distributions of power law tails:

Consider a set of N samples, randomly chosen according to the probability distribution $P_k = ck^{-\gamma}$ where k = 1,2,3,....
(a) Estimate min k_{max}, the approximate minimum of the largest sample in the system, finding how it depends on N.
(Hint: we expect on the order of 1 of the N samples to have a value of min k_{max} or greater.)
Hint - Some visual help on setting this problem up:

http://www.youtube.com/watch?v=4tqlEuXA7QQ

(b) Determine the average value of samples with value $k \geq \min k_{max}$ to find how the expected value of k_{max} (i.e., $\langle k_{max} \rangle$) scales with N.

Notes:

- For language, this scaling is known as Heap's law.
- In a later assignment, we will test this scaling by (thoughtfully) sampling form power-law size distributions

Responses:
(a) Since at any given point, the maximum in a given sample is given by the probability that 1 in N samples is greater than the rest, with value k^*, we have that

$$\frac{1}{N} = \int_{k^*}^{\infty} cx^{-\gamma} dx$$
$$\frac{1}{N} = \frac{c}{-\gamma+1} x^{-\gamma+1}|_{k^*}^{\infty}$$
$$\frac{1}{N} = 0 - \frac{c}{-\gamma+1} k^{*-\gamma+1}$$
$$N^{-1*(\gamma-1)} = -c^* k^*$$
$$k^* \propto N^{\frac{1}{\gamma-1}}$$

(b) To find the average, we find the first moment of the distribution evaluated from k^* to infinity.

$$\langle x \rangle = N \int_{k^*}^{\infty} xcx^{-\gamma} dx$$
$$\langle x \rangle = N \int_{k^*}^{\infty} cx^{-\gamma+1} dx$$
$$\langle x \rangle = N \frac{c}{-\gamma+2} x^{-\gamma+2}|_{k^*}^{\infty}$$

- For $1 < \gamma \leq 2$, we have infinite mean
- For $\gamma > 2$, we have:

$$\langle k^* \rangle = 0 - \frac{cN}{-\gamma+2} k^{*-\gamma+2}$$
$$\langle k^* \rangle = -\frac{cN}{-\gamma+2} k^{*-\gamma+1} k^*$$
$$\langle k^* \rangle = c^* \frac{c_2^* 1}{N} k^*$$
$$\langle k^* \rangle = c_3^* N^{\frac{1}{\gamma-1}}$$
$$N^{\frac{1}{\gamma-1}} \propto \langle k^* \rangle$$

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Create an allotaxonomograph
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
datevec1 = [2002];

if (~exist('elements'))
    textfile1 = sprintf('/data/storywrangler/babynames/names-boys2002.csv');

    textfile2 = sprintf('/data/storywrangler/babynames/names-girls2002.csv');

    story_wrangler_twitter_data(1).table = readtable(textfile1,...
                                                     'filetype','text',...
                                                     'delimiter',',');
    story_wrangler_twitter_data(2).table = readtable(textfile2,...
                                                     'filetype','text',...
                                                     'delimiter',',');

    %% subsample for latin characters
    for i=1:length(story_wrangler_twitter_data)
        %%    indices =
        %%    ~cellfun(@isempty,regexp(story_wrangler_twitter_data(i).table{:,1},'^[@#-''A-Za-
z]+$','match'));
        indices = ~cellfun(@isempty,regexp(story_wrangler_twitter_data(i).table{:,1},'^[A-Za-
z][-''A-Za-z]+$','match'));

        story_wrangler_twitter_data(i).table = ...
            story_wrangler_twitter_data(i).table(indices,:);
        length(indices)
    end

    indices = [1 2];
    for i=1:2
        elements(i).types = story_wrangler_twitter_data(indices(i)).table{:,1};
        elements(i).counts = story_wrangler_twitter_data(indices(i)).table{:,3};
%         elements(i).probs = story_wrangler_twitter_data(indices(i)).table{:,6};

        elements(i).ranks = tiedrank(-elements(i).counts);
        elements(i).totalunique = length(elements(i).types);
    end

    mixedelements = combine_distributions(elements(1),elements(2));
else
    fprintf(1,'elements not reloaded; delete if needed\n');
end

%% some settings
datetag_str = sprintf('Boys v Girls names in %04d', ...
                     datevec1(1));

settings.system1_name = sprintf('Boys names in %04d',...
                   datevec1(1));

settings.system2_name = sprintf('Girls names in %04d',...
                   datevec1(1));


settings.typename = 'word';

%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% general settings
%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
settings.binwidth = 0.21;
settings.topNshuffling = 25;
settings.topNshift = 40;
settings.topNdeltasum = 'all';

settings.max_plot_string_length = 15;
settings.max_shift_string_length = 25;

settings.imageformat.open = 'no';

%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% make some flip books
%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% example flipbook series for rank divergence for 1 and 2
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

settings.plotkind = 'rank';
%% settings.plotkind = 'probability';
%% settings.plotkind = 'count';

settings.instrument = 'rank divergence';
%% settings.instrument = 'probability divergence';
%% settings.instrument = 'alpha divergence type 2';

%% move the shift (adds to 0.60)
settings.xoffset = +0.05;

%% alphavals = [(0:18)/12, 2, 3, 5, 10, Inf]';

settings.alpha = Inf;

tag = sprintf('%04d',...
              datevec1(1));

figallotaxonometer9000(mixedelements,tag,settings);
```