# Similarity to a single set

Lee Naish
Computing and Information Systems
University of Melbourne

(a work in progress)

# Outline

The problem

Set similarity measures

Properties of set similarity measures

- scaling
- monotonicity
- symmetries

Relative "weight" of matches versus non-matches

Choosing a similarity measure

An experiment

Conclusion

# The problem

Suppose we have a finite number of cases, some subset of which have a particular attribute, the *base attribute*

There are also a number of other attributes each case may have

The problem is to rank all these attributes according to how well they "correlate" with the base attribute

Equivalently, we say the *base set* $B$ is the set of cases with the base attribute, etc, and we rank sets according to how "similar" they are to the base set

This "STASS" problem is pervasive in science. For example, we might want to know *why* certain things have some attribute. Finding a correlation can be the first step in determining a cause.

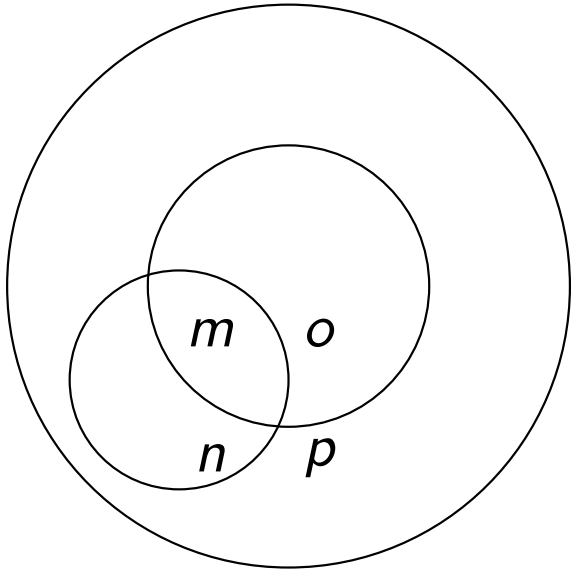
It can be generalised to similarity of arbitrary pairs of sets, correlation of non-binary attributes, clustering, interestingness of association rules, ...

# The problem

More terminology: $M = |B|$ and $N = |\bar{B}|$

Comparison of sets $B$ and $A$ can be viewed as follows:

| | $A$ | $\bar{A}$ | Total |
|---|---|---|---|
| $B$ | $m$ | $o$ | $M$ |
| $\bar{B}$ | $n$ | $p$ | $N$ |




Set similarity can be measured by a numeric function $f(m, n, o, p)$ and the results used to rank the sets/attributes

Because the base set fixes $M$ and $N$ we use these as parameters instead of $o$ and $p$, and write $f_N^M(m, n)$

# Example: program spectra for debugging

Estimate the relative likelihood of statements being buggy by correlating execution of statements in a test case with failure of the test case

$M$ is the number of failed tests, $N$ the number of passed tests, $m$ the number of failed tests which execute a particular statement and $n$ the number of passed tests which execute the statement

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $m$ | $n$ |
|---|---|---|---|---|---|---|---|
| $S_1$ | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| $S_2$ | 1 | 1 | 0 | 1 | 0 | 2 | 1 |
| $S_3$ | 1 | 1 | 1 | 0 | 1 | 2 | 2 |
| ⋮ | | | | | | | |
| Base | 1 | 1 | 0 | 0 | 0 | $M = 2$ | $N = 3$ |

# Set similarity measures

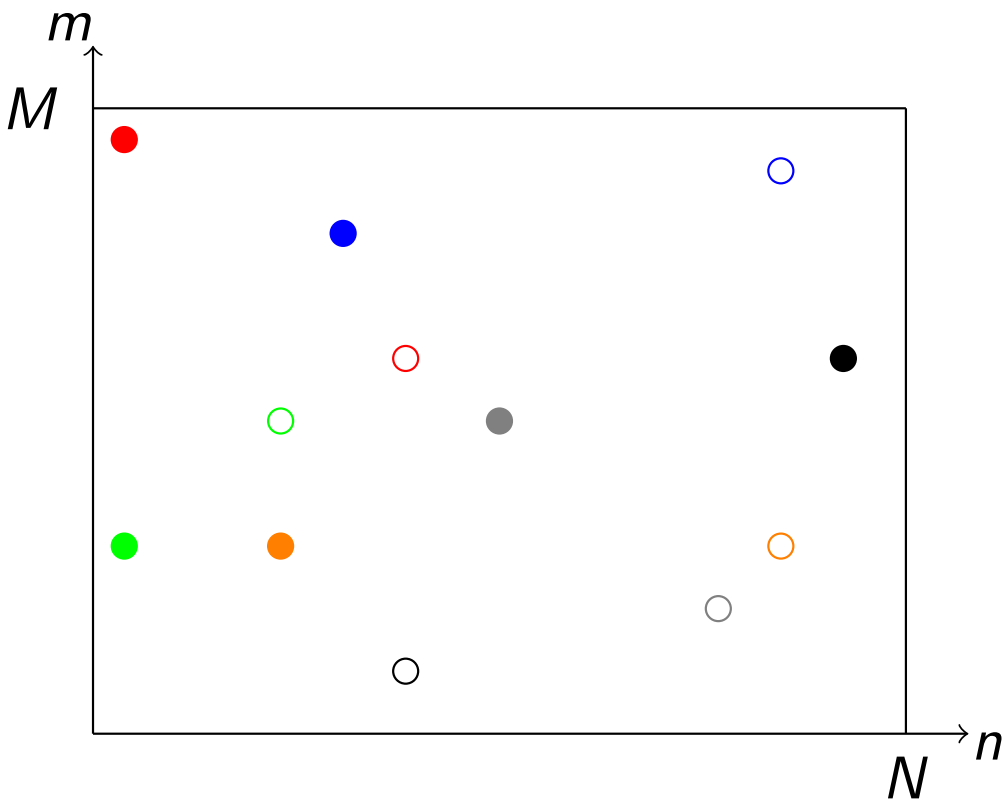
*Many* similarity measures have been used, in a wide variety of domains

| Name | Formula | Name | Formula |
|---|---|---|---|
| Jaccard | $\frac{m}{M+n}$ | Tarantula | $\frac{\frac{m}{M}}{\frac{m}{M}+\frac{n}{N}}$ |
| Russell | $\frac{m}{M+N}$ | Zoltar | $\frac{m}{M+n+\frac{10000(M-m)n}{m}}$ |
| Rogers | $\frac{m+N-n}{M+N}$ | Cosine | $\frac{m}{\sqrt{M(m+n)}}$ |
| Faith | $\frac{m+\frac{1}{2}(N-n)}{M+N}$ | Pearson | $\frac{Nm-Mn}{\sqrt{MN(m+n)(M+N-m-n)}}$ |
| Ample2 | $\frac{m}{M}-\frac{n}{N}$ | Ample | $\left\lvert\frac{m}{M}-\frac{n}{N}\right\rvert$ |
| *Op* | $m-\frac{n}{N+1}$ | Added Value | $\frac{m}{\max(M,m+N-n)}$ |
| Wong3 | $m-h,\ \text{where } h=\begin{cases} n & \text{if } n\leq 2 \\ 2+0.1(n-2) & \text{if } 2<n\leq 10 \\ 2.8+0.001(n-10) & \text{if } p>10 \end{cases}$ | | |

What is the best one for *my* problem?

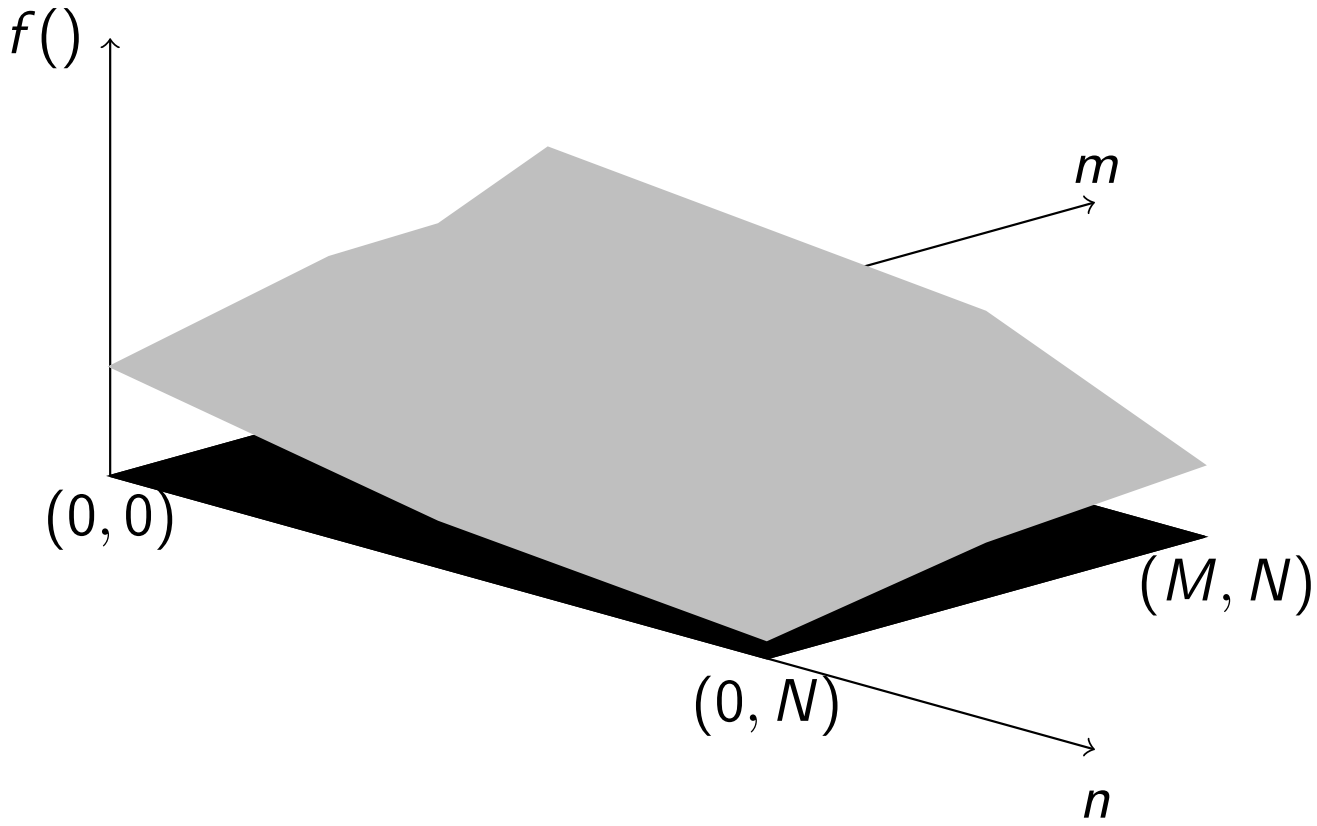
# A geometrical view

A STASS problem can be viewed a collection of points in a rectangular domain, which we need to rank



# Set similarity measures

Set similarity measures can be viewed as a surface over the domain



# Properties of set similarity measures

There are many properties of set similarity measures which seem desirable, eg, for philosophical reasons

Some of these properties are incompatible

Some of them are overly restrictive, at least for the STASS problem

For example, we only care about the relative value of two measures (which point is ranked higher), not the "absolute" value of a measure

We use the *result of comparison* of two numbers $x$ and $y$:
$C(x, y)$, is 1 if $x > y$, 0 if $x = y$ and -1 if $x < y$

Also, we only care about relative values where $M$ and $N$ the same for the two points

# Uniform scalability

We would expect that the ranking would be preserved if $M$, $N$, $m$ and $n$ were all multiplied by some scaling factor $s$

Eg, a measure $f$ is *absolute uniform scalable* if

$$f_N^M(m,n) = f_{sN}^{sM}(sm,sn)$$

For STASS we propose a "relative" rather than "absolute" definition

A measure $f$ is *uniform scalable* if

$$C(f_N^M(m,n), f_N^M(m',n')) = C(f_{sN}^{sM}(sm,sn), f_{sN}^{sM}(sm',sn'))$$

# General scalability, Null invariance

Often the cases with the base attribute are collected separately from those without the base attribute. The relationship between $M$ and $N$ is determined by the experimental design rather than the underlying domain

A measure $f$ is *general scalable* if

$$C(f_N^M(m,n), f_N^M(m',n')) = C(f_{tN}^{sM}(sm,tn), f_{tN}^{sM}(sm',tn'))$$

Adding extra cases which don't have the base attribute or any of the other attributes (arguably) should not affect the ranking

A measure $f$ is *null-invariant* if for all points where $f$ is defined and all $k$

$$C(f_N^M(m,n), f_N^M(m',n')) = C(f_{N+k}^M(m,n+k), f_{N+k}^M(m',n'+k))$$

Other papers define "absolute" versions of these properties

# Monotonicity

A measure $f$ is *monotone* if it is monotonically increasing in $m$ and monotonically decreasing in $n$

$$C(m, m') = C(f_N^M(m, n), f_N^M(m', n))$$

$$C(n, n') = -C(f_N^M(m, n), f_N^M(m, n'))$$

For many measures these conditions hold for most of the domain, but not when $m = 0$, for example

They can often be tweeked to make them monotone, eg Jaccard-m can be defined as $(m + \epsilon)/(M + n)$, or we can add $\epsilon$ only when $m = 0$

**Proposition**: No measure is absolute general scalable, absolute null-invariant and monotone (but our weaker definitions allow it, eg, $Op$)

# Symmetries

There are 16 different reflections/rotations (basic symmetries) of a surface. We can invert the surface. We can swap $n$ with $N-n$. We can swap $m$ with $M-m$. We can swap $m$ with $n$ and $M$ with $N$.

Only three such transformations preserve monotonicity

Reflection in $m=n$ and inversion: $f$ is *nm-antisymmetric* if

$$C(f_N^M(m,n), f_N^M(m',n')) = -C(f_M^N(n,m), f_M^N(n',m'))$$

$180^\circ$ rotation and inversion: $f$ is *$\bar{m}\bar{n}$-antisymmetric* if

$$C(f_N^M(m,n), f_N^M(m',n')) = -C(f_N^M(M-m,N-n), f_N^M(M-m',N-n'))$$

Reflection in $m=n$ and $180^\circ$ rotation: $f$ is *$\bar{n}\bar{m}$-symmetric* if

$$C(f_N^M(m,n), f_N^M(m',n')) = C(f_M^N(N-n,M-m), f_M^N(N-n',M-m'))$$

We can define three duals of a surface/similarity measure, eg $\mathcal{D}^{\bar{n}\bar{m}}(f)$

# Correlation antisymmetry and weight symmetry

The line $m = n$ is important for the basic symmetries, but the line of statistical independence is $Nm = Mn$, between $(0, 0)$ and $(M, N)$

Points above this line have positive correlation and points below have negative correlation; we may have anti-symmetry around this line

$f$ is *nm-scaled-antisymmetric*, or *correlation-antisymmetric*, if

$$C(f_N^M(m, n), f_N^M(m', n')) = -C(f_N^M(Mn/N, Nm/M), f_N^M(Mn'/N, Nm'/M))$$

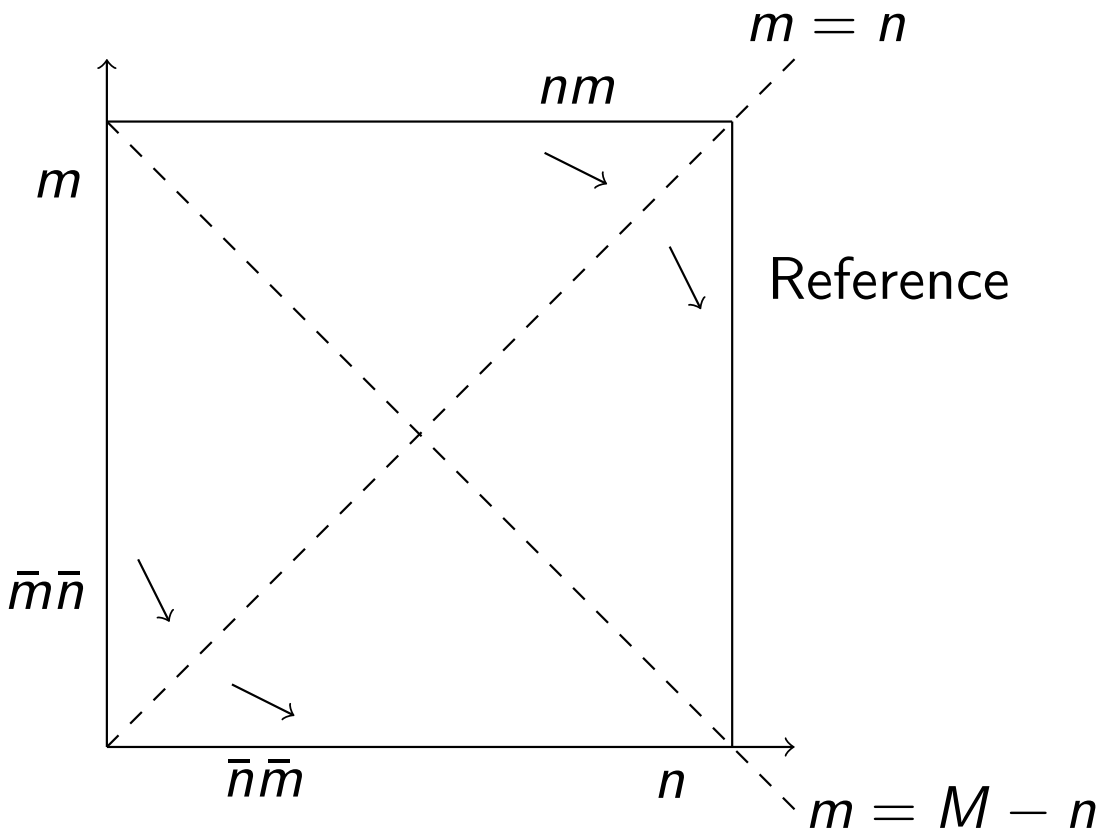
(We generalise to non-integers)

Similarly, we may have symmetry around the line $Mn = MN - Nm$, between points $(M, 0)$ and $(0, N)$

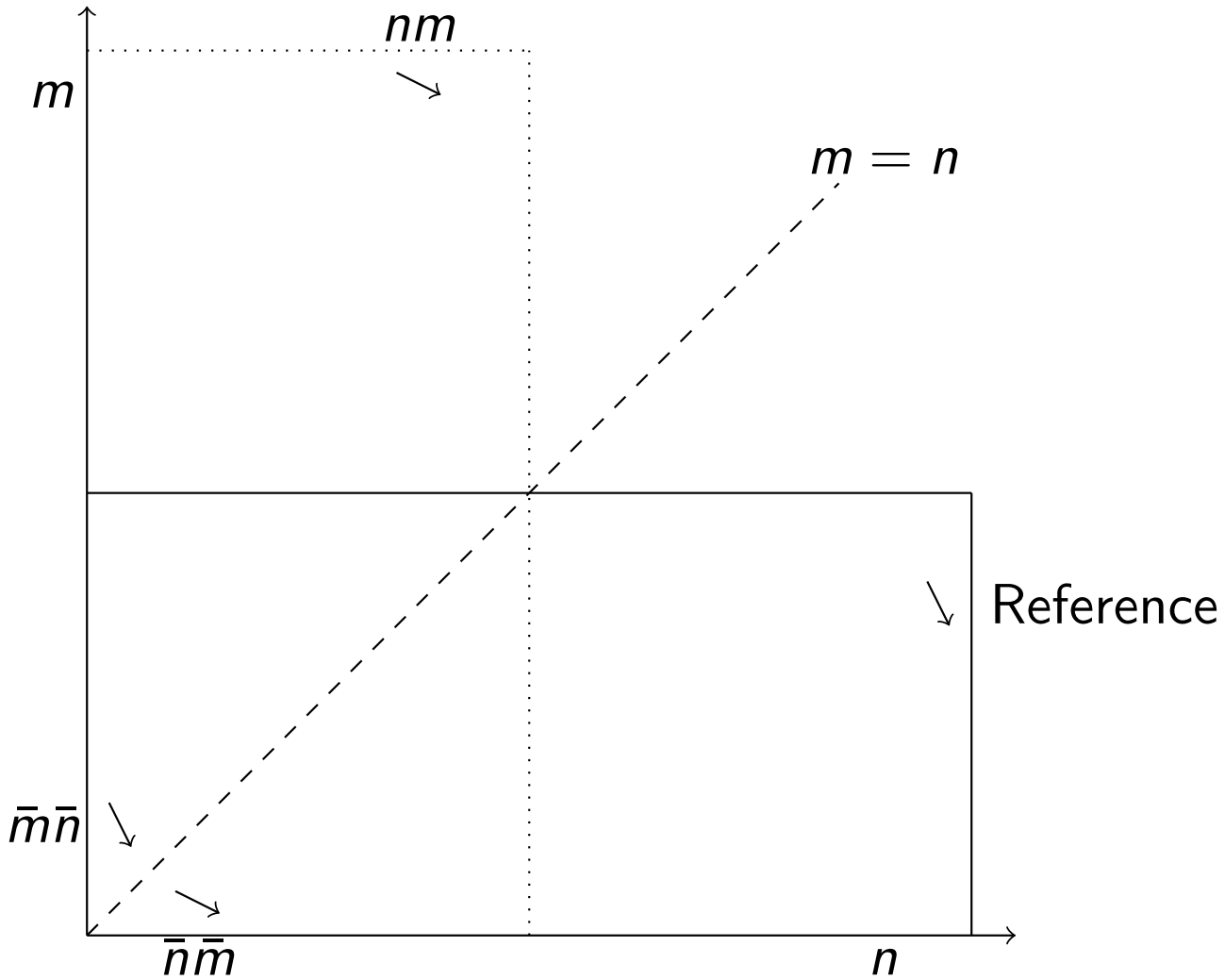
A measure $f$ is *$\bar{n}\bar{m}$-scaled-symmetric*, or *weight-symmetric*, if

$$C(f_N^M(m, n), f_N^M(m', n')) =$$
$$C(f_N^M(M - Mn/N, N - Nm/M), f_N^M(M - Mn'/N, N - Nm'/M))$$

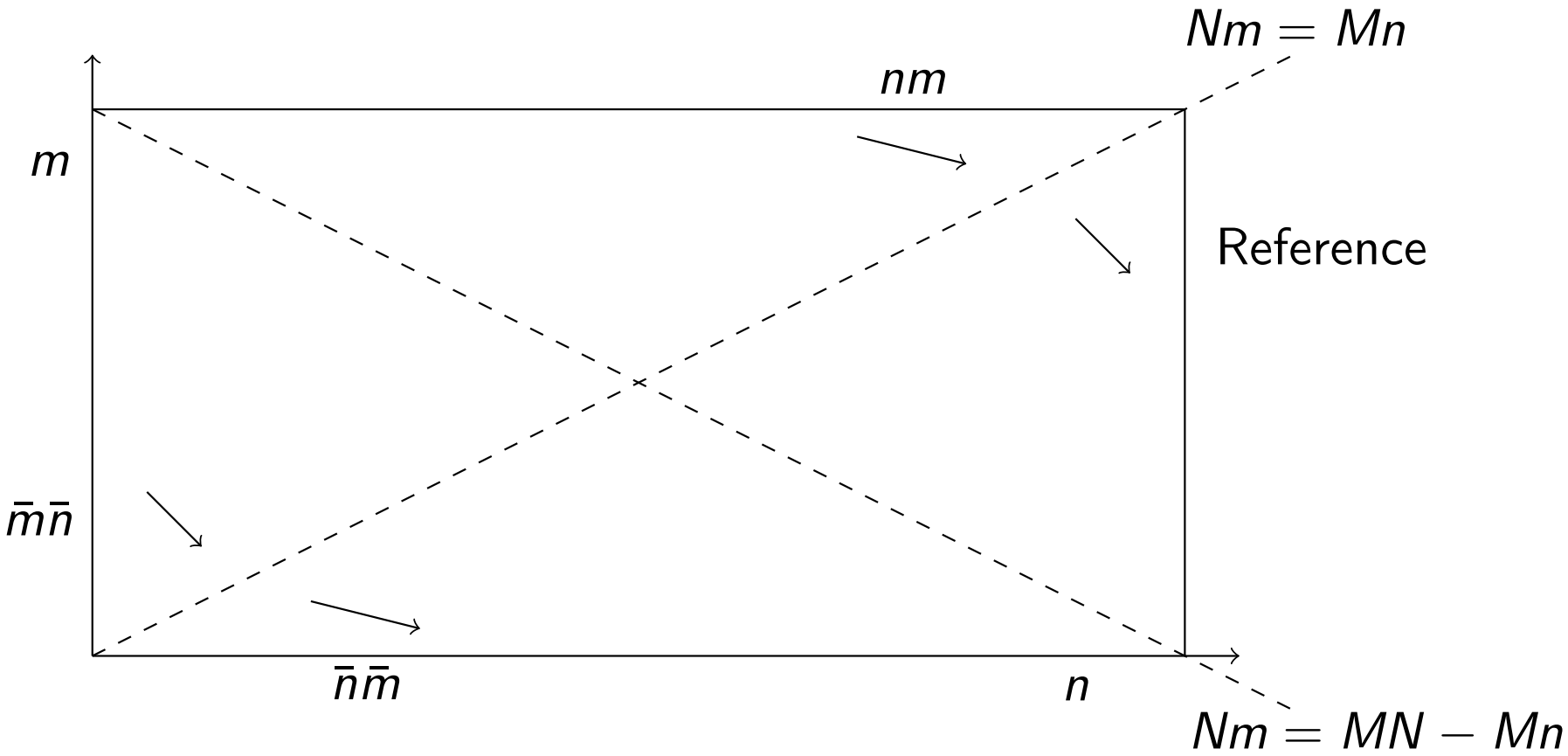
# Basic symmetries, $M = N$



# Basic symmetries, $M \neq N$



# Scaled symmetries, $M \neq N$



# Relative weight of $m$ and $n$

Many measures place more importance or *weight* on matches ($m$) than non-matches ($n$)

For Jaccard, $m/(M+n)$, doubling $m$ requires more than doubling $n$ to compensate ($n$ has less weight)

$f \sqsupseteq_N^M g$ ($f$ has greater or equal *set m-weight* than $g$ for $M$ and $N$) if

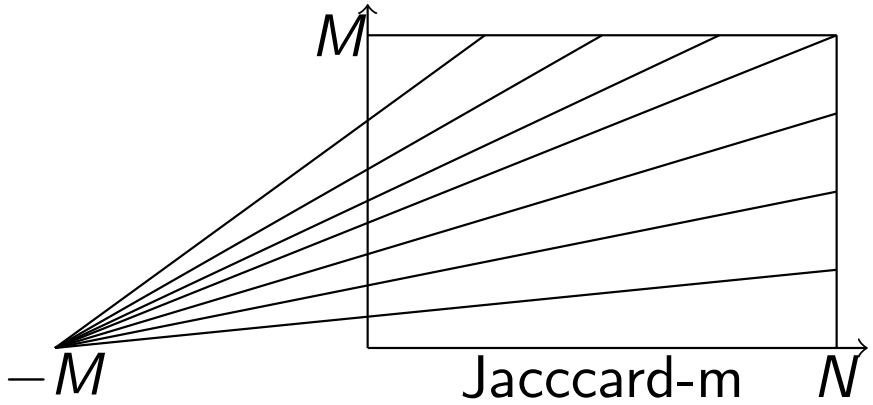
1. if $m \geq m'$ then $C(f_N^M(m,n), f_N^M(m',n')) \geq C(g_N^M(m,n), g_N^M(m',n'))$
2. if $m < m'$ then $C(f_N^M(m,n), f_N^M(m',n')) \leq C(g_N^M(m,n), g_N^M(m',n'))$

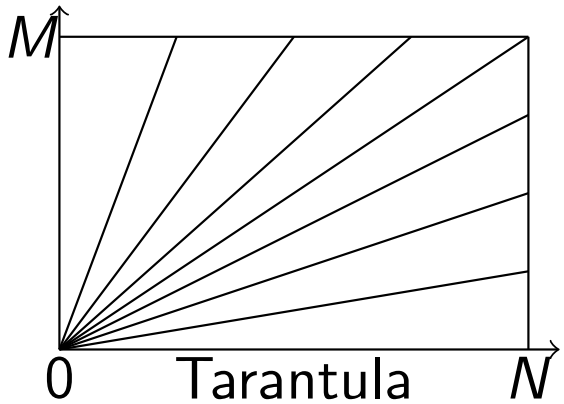
For given $M$ and $N$, monotone measures form a *complete lattice*

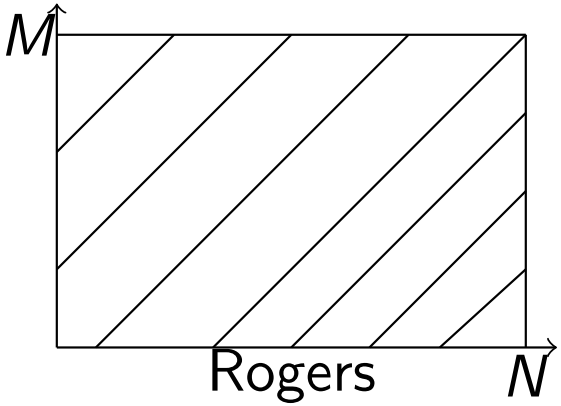
$Op$ is the top element (maximal set $m$-weight) and its weight-dual $\mathcal{D}^w(Op)$ is the bottom element

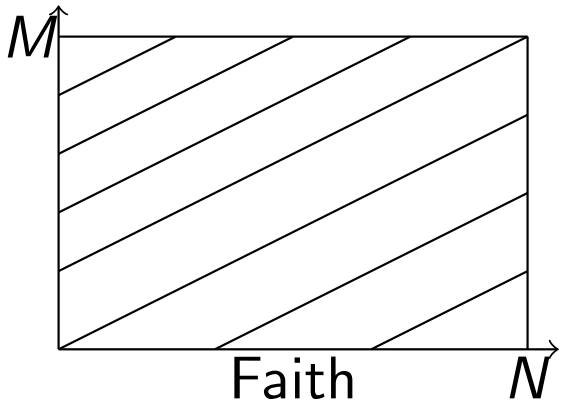
The lattice is symmetric, with $f \sqsupseteq_N^M g$ iff $\mathcal{D}^w(g) \sqsupseteq_N^M \mathcal{D}^w(f)$

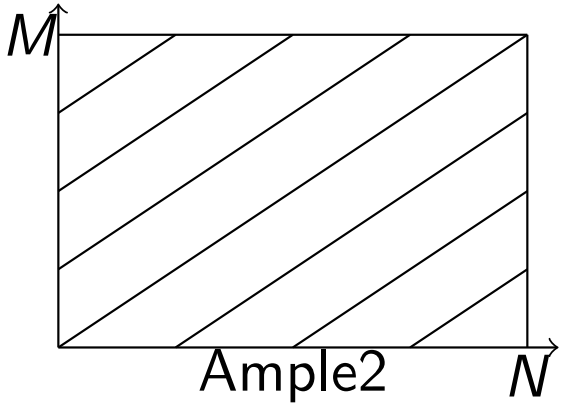
# Contours of some measures



# Quantifying the relative weight of $m$ and $n$

We can also quantify the relative $m$-weight by ranking all points in the domain and seeing how similar the ordering is to $Op$ and $\mathcal{D}^w(Op)$ (we scale the number of inversions so $Op$ is 1, $\mathcal{D}^w(Op)$ is 0 and weight-symmetric measures are 0.5)

| $M$ | $N$ | Kul2 | Cos | Zol | J | D-J | Tar |
|---|---|---|---|---|---|---|---|
| 10 | 50 | 0.79 | 0.72 | 0.78 | 0.65 | 0.35 | 0.50 |
| 30 | 30 | 0.79 | 0.81 | 0.78 | 0.82 | 0.18 | 0.50 |
| 50 | 10 | 0.89 | 0.93 | 0.77 | 0.96 | 0.04 | 0.50 |

Or we can just look at (eg) the top-most corner of the domain

| $M$ | $N$ | Kul2 | Cos | Zol | J | D-J | Tar |
|---|---|---|---|---|---|---|---|
| 10 | 50 | 0.33 | 0.22 | 0.49 | 0.16 | 0.84 | 0.03 |
| 30 | 30 | 0.65 | 0.66 | 0.59 | 0.67 | 0.33 | 0.03 |
| 50 | 10 | 0.93 | 0.94 | 0.62 | 0.95 | 0.05 | 0.04 |

# Using domain knowledge to choose a measure

From previous results we know $Op$ is the best possible measure for debugging programs with a single bug. We have used simple models of the debugging problem. Eg, there is a single bug which is executed half the time and causes failure half the time it is executed (on average)

From a model we can create the most likely contingency table for just the bugs (causes of the base attribute)

| | $A_c$ | $\bar{A}_c$ | Total |
|---|---|---|---|
| $B$ | $m_c$ | $o_c$ | $M_c$ |
| $\bar{B}$ | $n_c$ | $p_c$ | $N_c$ |

# Using domain knowledge to choose a measure

From previous results we know $Op$ is the best possible measure for debugging programs with a single bug. We have used simple models of the debugging problem. Eg, there is a single bug which is executed half the time and causes failure half the time it is executed (on average)

From a model we can create the most likely contingency table for just the bugs (causes of the base attribute)

| | $A_c$ | $\bar{A}_c$ | Total |
|---|---|---|---|
| $B$ | $m_c$ | $o_c$ | $M_c$ |
| $\bar{B}$ | $n_c$ | $p_c$ | $N_c$ |

| | $A_c$ | $\bar{A}_c$ | Total |
|---|---|---|---|
| $B$ | 25 | 0 | 25 |
| $\bar{B}$ | 25 | 50 | 75 |

# Using domain knowledge to choose a measure

From previous results we know $Op$ is the best possible measure for debugging programs with a single bug. We have used simple models of the debugging problem. Eg, there is a single bug which is executed half the time and causes failure half the time it is executed (on average)

From a model we can create the most likely contingency table for just the bugs (causes of the base attribute)

| | $A_c$ | $\bar{A}_c$ | Total |
|---|---|---|---|
| $B$ | $m_c$ | $o_c$ | $M_c$ |
| $\bar{B}$ | $n_c$ | $p_c$ | $N_c$ |

| | $A_c$ | $\bar{A}_c$ | Total |
|---|---|---|---|
| $B$ | 25 | 0 | 25 |
| $\bar{B}$ | 25 | 50 | 75 |

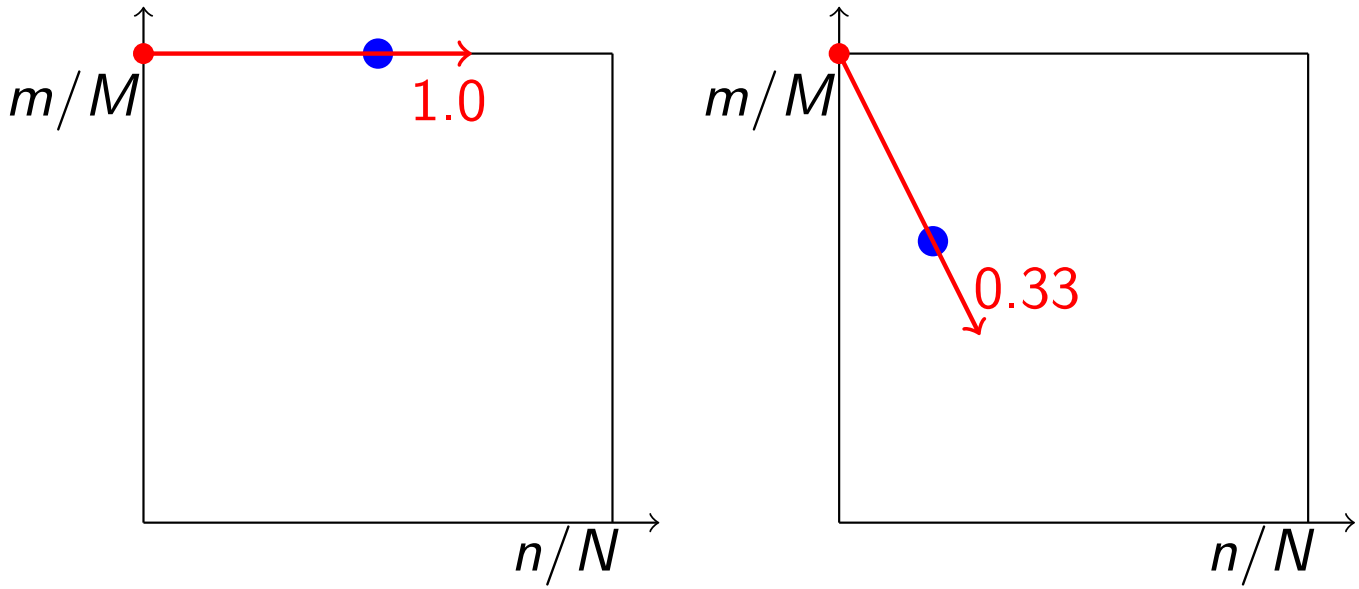
If there is a single bug, $o_c = 0$ and bugs always appear at the top edge of the domain ($m = M$)

Non-bugs are typically spread across the whole domain

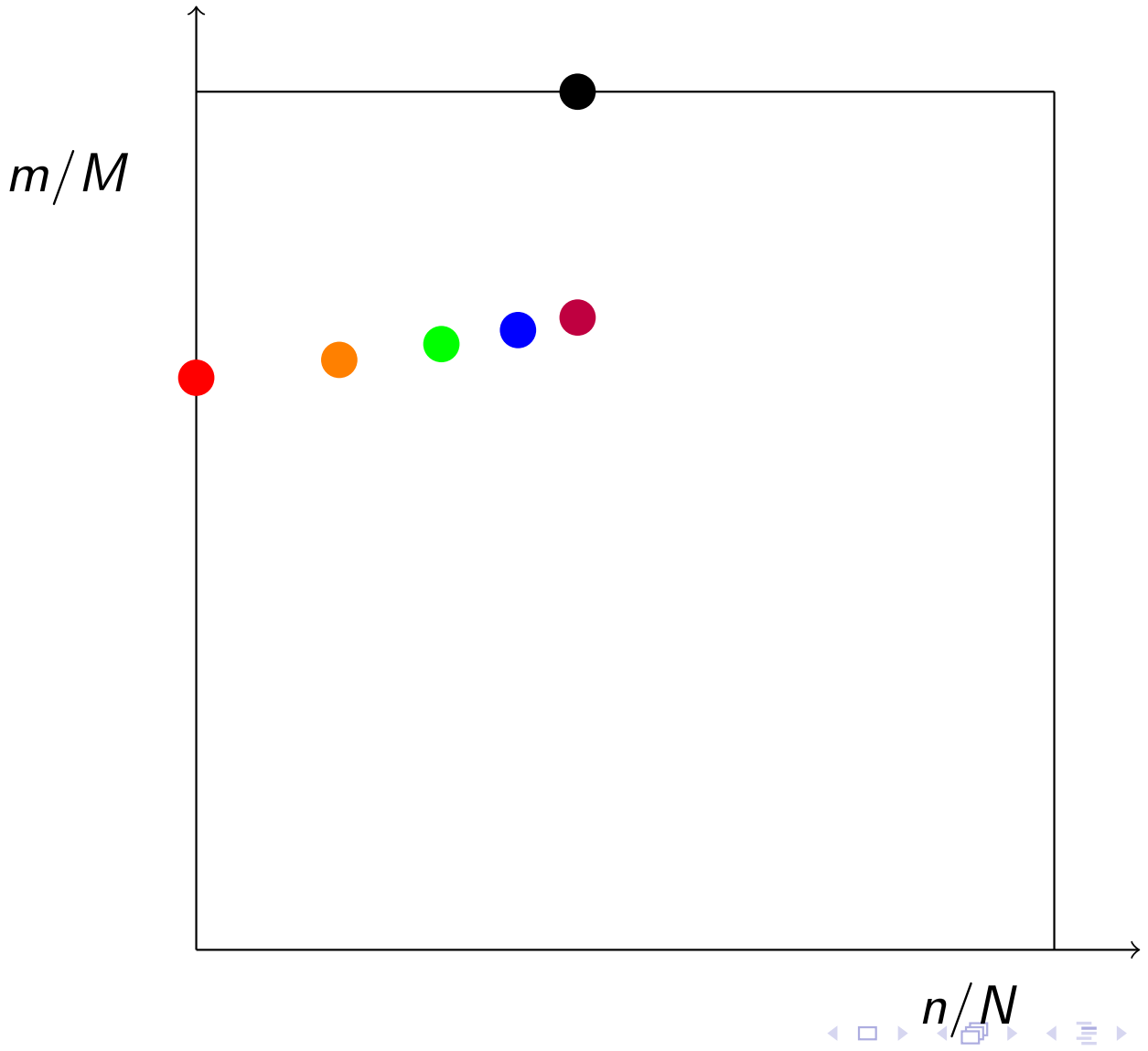
# Using domain knowledge to choose a measure

For programs with *deterministic* bugs (which cause failure whenever they are executed), $n_c = 0$, so bugs always appear at the left edge of the domain, and $\mathcal{D}^w(Op)$ is the best possible measure

From a contingency table for the causes, we can calculate the expected $m$ and $n$ values scaled to the range 0–1, $m_e = m_c/M_c$ and $n_e = n_c/N_c$, and the *ideal m-weight*: $n_e/(1 - m_e + n_e)$



# An experiment with varying ideal $m$-weights . . .



# Six simple debugging models

```
if ... then S1;
if ... then S2;
if ... then S3;
if ... then S4;
```

Each statement S1–S4 has a 50% chance of being executed and if it is buggy the test case has a certain chance of failing

We have experimented with a single bug model and five multiple bug models with the latter chance being 20%, 40%, 60%, 80% and 100%, respectively

The measures/surfaces we used were planes, varying from $Op$ to $\mathcal{D}^w(Op)$ ($cm/M - (1-c)n/N$ with $c$ varying from $1-\epsilon$ down to $\epsilon$)

# Experimental results

| | M1 | M2 | M3 | M4 | M5 | M6 |
|---|---|---|---|---|---|---|
| $m_c$, $o_c$ | 40, 0 | 56, 20 | 104, 40 | 144, 60 | 176, 80 | 200, 100 |
| $n_c$, $p_c$ | 160, 200 | 144, 180 | 96, 160 | 56, 140 | 24, 120 | 0, 100 |
| ideal | 1 | 0.63 | 0.57 | 0.49 | 0.34 | 0 |
| best plane | 1 | 0.75 | 0.53 | 0.47 | 0.40 | 0 |
| inversions | 1 | 0.85 | 0.56 | 0.44 | 0.34 | 0 |
| $Op$ | **89.14** | 86.19 | 89.17 | 90.62 | 91.22 | 91.25 |
| P75 | 89.14 | **86.19** | 89.25 | 91.17 | 92.87 | 94.85 |
| P53 | 88.94 | 86.08 | **89.50** | 92.17 | 94.65 | 97.44 |
| P47 | 88.58 | 85.87 | 89.34 | **92.20** | 94.85 | 97.80 |
| P40 | 88.00 | 85.59 | 89.00 | 92.02 | **94.88** | 97.93 |
| $\mathcal{D}^w(Op)$ | 73.87 | 81.51 | 86.51 | 90.87 | 94.60 | **98.02** |

# Further work

There can be multiple causes, with different expected values and distributions

There can be multiple non-causes

There can be various ways of measuring performance

Some insights into STASS should be applicable to other more general problems

# Conclusions

The STASS problem is pervasive but surprisingly poorly understood

- Monotonicity is important
- The scaled versions of symmetry are important
- We now have greater understanding of how domain knowledge can be used to choose and/or create measures which perform well