Towards a Synergy-based Approach to Measuring Information Modification

Joseph T. Lizier, Benjamin Flecker and Paul L. Williams
Publications


Measuring local information modification

- Distributed computation is often discussed in terms of information storage, transfer and modification; e.g. (Langton, 1990).
- We have rigorous measures for information storage and transfer and their dynamics in time and space (*Information Dynamics*).

→ We seek a rigorous measure for **information modification**.
→ We seek a **local measure** for the **dynamics** of modification.

- The **Partial Information Decomposition** (PID) approach shows promise for application – we explore how it could be applied to modification and whether this can be localised.
Contents

- Information theory background: local information measures;
- Information dynamics and information modification;
  - Requirements for a measure of information modification;
- The Partial information decomposition approach;
- Application of PID to local information modification:
  - New axiom for localisation;
  - $I_{\min}$ shown to not satisfy this.
Information-theoretic concepts: 1. Shannon entropy

\[ H(X) = - \sum_x p(x) \log_2 p(x) = \langle - \log_2 p(x) \rangle \]

\[ H(X|Y) = - \sum_{x,y} p(x,y) \log_2 p(x|y) \]

Demonstrated by Shannon (1948) as the unique formulation to satisfy:

1. **Continuity** w.r.t. \( p(x) \);
2. **Monotonic increase** with the number of equally-likely choices for \( x \);
3. **Grouping**: “If a choice (can) be broken down into two successive choices, the original \( H \) should be the weighted sum of the individual values of \( H \); i.e. \( H \) is independent of how the process is divided into parts.”
Information-theoretic concepts: 2. Mutual information (MI)

\[ I(X; Y) = H(X) + H(Y) - H(X, Y) \]

\[ = \sum_{x,y} p(x, y) \log_2 \frac{p(x|y)}{p(x)} \]

\[ = \langle \log_2 \frac{p(x|y)}{p(x)} \rangle \]

Venn diagram from (MacKay, 2003)
Information-theoretic concepts: 3. Conditional MI

\[ I(X; Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z) \]
\[ = \left\langle \log_2 \frac{p(x|y,z)}{p(x|z)} \right\rangle \]

\[ I(X; Y, Z) = I(X; Z) + I(X; Y|Z) \]

\( I(X; Y|Z) \) can be either **larger or smaller** than \( I(X; Y) \):

- Conditioning removes **redundant** information in \( Y \) and \( Z \) about \( X \);
- Conditioning includes **synergistic** information in the pair \( \{Y, Z\} \) about \( X \).

→ Can’t measure these effects separately with traditional information theory.
Information-theoretic concepts: 4. local measures

We can write local (or point-wise) information-theoretic measures for specific observations/configurations \( \{x, y, z\} \):

\[
\begin{align*}
    h(x) &= - \log_2 p(x), \\
    h(x|y) &= - \log_2 p(x|y), \\
    i(x; y) &= \log_2 \frac{p(x|y)}{p(x)}, \\
    i(x; y|z) &= \log_2 \frac{p(x|y, z)}{p(x|z)}.
\end{align*}
\]

- We have \( H(X) = \langle h(x) \rangle \) and \( I(X; Y) = \langle i(x; y) \rangle \), etc.
- If \( X, Y, Z \) are time-series, local values measure dynamics over time.
Information-theoretic concepts: 4. local measures

Q: Where do these local values come from, and what do they mean?

Local entropy: \( h(x) = -\log_2 p(x) \):

- \( h(x) \) is the uncertainty attributed to the specific symbol \( x \) or information required to uniquely specify/predict that symbol.
- Less probable outcomes \( x \) have higher information content.
- \( h(x) \geq 0 \)
- Can be derived as the unique form satisfying (Ash, 1965):
  - \( h(p_1 \times p_2) = h(p_1) + h(p_2) \);
  - monotonic decrease of \( h(p) \) with \( p \);
  - continuity with \( p \).
- \( h(x) \) is the code-length for symbol \( x \) in an optimal encoding scheme for measurements of \( X \).
Information-theoretic concepts: 4. local measures

Q: Where do these local values come from, and what do they mean?

Local mutual information: \( i(x; y) = \log_2 \frac{p(x|y)}{p(x)} : \)

- \( i(x; y) \) is the MI attributed to the specific symbol pair \( x, y. \)
- MI increases as \( p(x \mid y) \) becomes larger than \( p(x) . \)
- Local MI can be negative – where \( p(x \mid y) \) is lower than \( p(x) \), i.e. \( y \) was misinformative about \( x. \)

\[ i(x; y) = h(x) - h(x \mid y) : \text{coding penalty for } x \text{ in not being aware of } y (\text{under optimal encoding schemes for } X \text{ or } X \text{ given } Y) . \]

\( \rightarrow \) Fano (1961) set criteria to uniquely define local & cond’l MI:
- once-differentiability,
- similar form for conditional MI,
- additivity: \( i(\{y, z\}; x) = i(y; x) + i(z; x \mid y) , \) and
- separation for independent ensembles.
Credit assignment problem and information modification

**Fundamental question:** How can we describe the assignment of information in a target variable amongst several sources?

We will bring together two complementary approaches to study information modification:

1. **Information dynamics** (Lizier et al., 2008, 2010, 2012);
2. **Partial information decomposition** (Williams and Beer, 2010a,b).
Information dynamics

Studies computation of the next state of a target variable in terms of information storage, transfer and modification: (Lizier et al., 2008, 2010, 2012)

Active information storage:

\[ A_X = \langle a_X(n) \rangle = \langle i(x_{n+1}; x_n^{(k)}) \rangle \]

Information from past state that is in use in predicting the next value
Information dynamics

Studies computation of the next state of a target variable in terms of information storage, transfer and modification: (Lizier et al., 2008, 2010, 2012)

Transfer entropy:

\[ T_{Y \rightarrow X} = \langle t_{Y \rightarrow X}(n) \rangle = \langle i(x_{n+1}; y_n | x^{(k)}_n) \rangle \]

Info from source that helps to predict destination value in the context of destination’s past state.

(Higher order) Conditional transfer entropy:

\[ T_{Y_1 \rightarrow X | Y_2} = \langle t_{Y_1 \rightarrow X | Y_2}(n) \rangle = \langle i(x_{n+1}; y_{1,n} | x^{(k)}_n, y_{2,n}) \rangle \]

Total information:

\[ H(X) = A_X + T_{Y_1 \rightarrow X} + T_{Y_2 \rightarrow X | Y_1} \]
Information dynamics

Studies computation of the next state of a target variable in terms of information storage, transfer and modification: (Lizier et al., 2008, 2010, 2012)

Active information storage:
\[ A_X = \langle a_X(n) \rangle = \left\langle i(x_{n+1}; x_n^{(k)}) \right\rangle \]

Transfer entropy:
\[ T_{Y\rightarrow X} = \langle t_{Y\rightarrow X}(n) \rangle = \left\langle i(x_{n+1}; y_n|x_n^{(k)}) \right\rangle \]

Why this perspective? These are well-understood terms; they can be measured on any type of time-series; and computation is the language in which dynamics are often described (Langton, 1990).
**Information dynamics in cellular automata**

Local information storage and transfer confirm conjectures (Langton and others) regarding computational roles of blinkers and gliders.

![Raw CA](a) Raw CA  ![Storage dynamics](b) Storage dynamics  ![Transfer (to right) dynamics](c) Transfer (to right) dynamics

(Lizier et al., 2008-2012)

J.T. Lizier - Java Information Dynamics Toolkit (JIDT)
http://code.google.com/p/information-dynamics-toolkit/

→ But we lack a *satisfactory* measure for *information modification* - hypothesized to confirm glider collisions as modifications (Langton, 1990).
Information modification

Langton (1990): interactions between transmitted and/or stored information which result in a modification of one or the other. A dynamic combination / synthesis / non-trivial processing of information from two or more (storage or transfer) sources.

A measure of information modification $M_X$ should:

1. be a proper information-theoretic quantity;

2. examine the interaction between the information storage $X^{(k)}$ and causal transfer sources $Y \in \{Y_1, \ldots, Y_g\}$;

3. allow local measurement $m_X$ at specific observed configurations \( (x_{n+1}, x_n^{(k)}, y_{1,n}, \ldots, y_{g,n}) \);

4. be extendible to an arbitrary number of sources $g$.

→ Our previous suggestions (Flecker et al. (2011); Lizier et al. (2010)) don’t properly qualify.
Partial information decomposition (PID)

**Abstract** framework to measure arbitrary *Pl-terms*: redundancies, synergies and unique contributions from source variables to a target (Williams and Beer, 2010a,b).

E.g. Decompose MI from *two* sources – only have *three* info-theoretic primitives (marked), but have *four unknown Pl-terms*:

Redundancy: \( \{ M \} \{ Y \} \)
Unique information: \( \{ M \} \) and \( \{ Y \} \)
Synergy: \( \{ M, Y \} \)

Conditional MI: \( I(X; Y|M) \)
PI-diagram for three source variables

More complicated - 7 primitives and 17 unknown PI-terms!
Partial information decomposition (PID)

\[ I(X; Y) \]
\[ I(X; M) \]
\[ I(X; M, Y) \]

Key: measure redundancy \( I_n(X; \{M\}, \{Y\}) \) and other PI-terms \( I_\partial \) follow via inclusion-exclusion algebra, if \( I_n(\mathbf{A}_1 \ldots \mathbf{A}_{r-1}, \mathbf{A}_r) \) conforms to a specific set of axioms:

**Axiom 1.** Symmetry: \( I_n \) is symmetric in the \( \mathbf{A}_i \)'s.

**Axiom 2.** Self-redundancy: \( I_n(X; \mathbf{A}_i) = I(X; \mathbf{A}_i) \).

**Axiom 3.** Monotonicity: \( I_n(X; \mathbf{A}_1 \ldots \mathbf{A}_{r-1}, \mathbf{A}_r) \leq I_n(X; \mathbf{A}_1, \ldots, \mathbf{A}_{r-1}) \) with equality if \( \mathbf{A}_{r-1} \subseteq \mathbf{A}_r \).
$I(X; M) \rightarrow A_X$

$I(X; Y_1|M) \rightarrow T_{Y_1 \rightarrow X}$

$I(X; Y_2|M, Y_1) \rightarrow T_{Y_2 \rightarrow X|Y_1}$

Flecker et al. (2011)
Previous approaches to info modification don’t qualify

1. Separable information (Lizier et al., 2010)
   \[ s = a_X + t_{Y_1 \rightarrow X} + t_{Y_2 \rightarrow X} < 0 \]

2. Highest order synergy (Flecker et al., 2011)
Identifying information modification in PI-diagram

- **White area**: information about $X$ that can be found in any one source $\{M, Y_1, Y_2\}$
- **Recall**: information modification as synthesis of information from two or more (storage or transfer) sources – **synergies**
- **Identify information modification in green and blue areas**: information about $X$ that can only be found in a pair or larger combination of sources.

Satisfies our requirements to measure information modification, if we have a localisable redundancy measure $I \cap \ldots$
Localising redundancy $I_{\cap}$

We propose a new axiom for a redundancy measure $I_{\cap}$ to be localisable:

**Axiom 5.** Localizability: There exists a local measure $i_{\cap}(x; a_1, \ldots, a_r)$ for the redundancy of a specific observation $\{x, a_1, \ldots, a_r\}$ of $\{X, A_1, \ldots, A_r\}$ such that:

1. $i_{\cap}(x; a_1, \ldots, a_r)$ satisfies the corresponding symmetry and self-redundancy axioms as per $I_{\cap}(X; A_1, \ldots, A_r)$;
2. $I_{\cap}(X; A_1, \ldots, A_r) = \langle i_{\cap}(x; a_1, \ldots, a_r) \rangle$;
3. $i_{\cap}(x; a_1, \ldots, a_r)$ is once-differentiable with respect to changes in $p(x, a_1, \ldots, a_r)$; and
4. $i_{\cap}(x; a_1, \ldots, a_r)$ is uniquely defined for the given candidate redundancy measure.
Localising redundancy $I_\cap$

**Axiom 5.** Localizability: *There exists a local measure* $i_\cap(x; a_1, \ldots, a_r)$ *for the redundancy of a specific observation* $\{x, a_1, \ldots, a_r\}$ *of* $\{X, A_1, \ldots, A_r\}$

- Has similar requirements to $I_\cap$ and local MI, but no requirement for $i_\cap$ to satisfy monotonicity – local MI values can increase or decrease with number of variables so long as average increases;
- Since local MI can be negative, so too can $i_\cap$;
- Sliding window methods are not local values;
- Motivation for a local redundancy measure goes beyond application for information modification: it would make any PI-term measurable on a local scale.
$I_{\text{min}}$ – a concrete measure for redundancy $I_n$

**Interpretation:** $I_{\text{min}}$ measures the *minimum* amount of information that can be found in any single source about the value of the target variable (averaged over all target values).

**Mathematical definition:** (Williams and Beer, 2010a)

$$I_{\text{min}}(X; A_1, \ldots, A_r) = \sum_s p(s) \min_{A_j} I(X = x; A_j),$$

$$I(X = x; A) = \sum_a p(a|x) \left[ \log_2 \frac{1}{p(x)} - \log_2 \frac{1}{p(x|a)} \right].$$
$I_{\text{min}}$ – a concrete measure for redundancy $I_{\cap}$

**Example 1:** OR function $X = A_1 + A_2$: $I_{\text{min}} = 0.311$ bits – e.g. $A_1$ and $A_2$ contain redundant information about the $x = 0$ outcome.

**Example 2:** *Two-bit copy problem* $X = \{A_1, A_2\}$:

$$I_{\text{min}}(\{A_1, A_2\}; A_1, A_2) = 1 \text{ bit}$$

Naive expectation: 1 bit of unique information for each variable, but actually get 1 bit of redundancy and 1 bit of synergy.

$\rightarrow I_{\text{min}}$ measures minimum information found in single sources, but does not specifically require each source to hold the *same* information.
New axiom and measures

To address the two-bit copy problem, a new axiom was proposed for $I_n$ (Harder et al., 2012):

**Axiom 4. Identity:** $I_n(A_1, A_2; A_1, A_2) = I(A_1; A_2)$.

New candidates have been suggested which satisfy Axiom 4:

- Harder et al. (2012) – information geometric method to compute distance between distributions.
Localising candidate measure $I_{\text{min}}$

**Intuition:** $i_{\text{min}}$ measures local MI from source which provided the minimum amount of information for the given target value.

**Mathematical definition:**

$$i_{\text{min}}(x; a_1, \ldots, a_r) = i(x; a_j) = \log_2 \frac{p(x \mid a_j)}{p(x)},$$

$$A_j = \arg \min_{A_j} I(X = x; A_j).$$

This is a unique form, since $I_{\text{min}}(X; A_1, \ldots, A_r) = I(X; A_j)$ for $A_j$ defined above, and $i_{\text{min}}(x; a_1, \ldots, a_r)$ must average to this.
Localising candidate measure $I_{min}$ - OR example

**OR logic gate:** $X = A_1 + A_2$

Redundancy $I_{min}(X; A_1, A_2) = 0.311$ bits.

**Local** redundancy $I_{min}(x; a_1, a_2)$ for each *almost* equiprobable configuration $(a_1, a_2)$:

<table>
<thead>
<tr>
<th>$a_1, a_2$</th>
<th>$x$</th>
<th>$p(a_1, a_2)$</th>
<th>$\delta \to 0^+$</th>
<th>$\delta \to 0^-$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1, a_2$</td>
<td>$x$</td>
<td>$p(a_1, a_2)$</td>
<td>arg min $I(X = x; A_j)$</td>
<td>$i(x; a_j)$</td>
</tr>
<tr>
<td>0,0</td>
<td>0</td>
<td>0.25</td>
<td>$A_1$</td>
<td>1</td>
</tr>
<tr>
<td>0,1</td>
<td>1</td>
<td>0.25 + $\delta$</td>
<td>$A_1$</td>
<td>-0.585</td>
</tr>
<tr>
<td>1,0</td>
<td>1</td>
<td>0.25 - $\delta$</td>
<td>$A_1$</td>
<td>0.415</td>
</tr>
<tr>
<td>1,1</td>
<td>1</td>
<td>0.25</td>
<td>$A_1$</td>
<td>0.415</td>
</tr>
</tbody>
</table>

→ $i_{min}$ is not continuous, nor unique; $I_{min}$ cannot be localised.
→ cannot use $I_{min}$ to measure local information modification.
Prospects with other measures – Harder et al. (2012)

\[ I_{\text{red}}(Z; X, Y) := \min \{ I_{\pi}^Z(X \downarrow Y), \]
\[ I_{\pi}^Z(Y \downarrow Y) \} \]

However:

- The projection is not guaranteed to be unique (and it is the projection that would determine the local values);
- the measure is not extendible to arbitrary number of sources;
Prospects with other measures – Griffith and Koch (2012)

\[ I_U(\{A_1, \ldots, A_n\}; X) = \min_{p(x'|x)} I(\{A_1, \ldots, A_n\}; X') \]

subject to: \( \{A_1, \ldots, A_n\} \rightarrow X \rightarrow X' \)

\[ I(A_i; X') = I(A_i; X) \quad \forall i \]

Again however, this maps to a non-unique PDF for computing local values.
Prospects with other measures

Final comments:
→ There may be scope to extend these measures in future, bearing our new axiom in mind.
→ Or, perhaps localizability cannot co-exist with the other axioms, as shown by Bertschinger et al. (2012) regarding strong symmetry and the existing axioms . . .
Conclusion and Future prospects

Contribution:
1. Linked information dynamics & PID to define info modification;
2. Additional localisability axiom for PID’s redundancy $I_\cap$;
3. Showed that $I_{\text{min}}$ is unsuitable for these.
4. Open-source PID/$I_{\text{min}}$ code.

Hopefully new redundancy measures will satisfy localizability ...
References


Thank You

CSIRO ICT Centre

Joseph Lizier

t  +61 2 9372 4711
e  Joseph.Lizier@csiro.au
w  http://lizier.me/joseph/

JL thanks the Max Planck Insitute for Mathematics in the Sciences for supporting this visit