# A NEW LOOK AT THE INFORMATIONAL GAIN OF SOFT DECISIONS

*Michael A. Lexa and Don H. Johnson*

Rice University
Department of Electrical and Computer Engineering
Houston, TX 77005-1892
amlexa@rice.edu, dhj@rice.edu

**ABSTRACT**

This paper develops a new systematic method of studying the benefits of 2-bit soft decisions by applying the concepts of information processing theory. We quantify performance in terms of the information transfer ratio and demonstrate the performance gain over hard decision detectors in several noise environments. In addition, we show that likelihood ratio tests maximize the information transfer ratio, and we propose a method of optimizing threshold values for the 2-bit soft decision detector.

## 1. INTRODUCTION

In our theory of information processing, information is defined only with respect to the ultimate receiver. Consequently, no single objective measure can quantify the information a signal expresses. For example, this paper (presumably) means more to a signal processing researcher than it does to a Shakespearean scholar. To probe how well systems process information, we resort to calculating how well an informational change at the input is expressed in the output. The complete theoretical basis of this theory can be found elsewhere [1]. Briefly, information is represented by the abstract quantity $\alpha$ and signals (here binary data) represent information. To quantify an informational change $\alpha_1 \rightarrow \alpha_2$, we calculate the information-theoretic distance, specifically the Kullback-Leibler distance[1](KL), between the probability distributions characterizing the *signals* that encode two pieces of information. We assume the signals, but not the information, are stochastic. The Data Processing Theorem (DPT) [1] says that the KL distance between the outputs of any system responding to the two inputs must be less than or equal to the distance calculated at the input. Here, we use this framework to quantify the informational gain achieved by 2-bit soft decision detectors over hard decision detectors.

We adopt the digital communication system model shown schematically in Figure 1. The input binary data word $\mathbf{u}_\alpha$ of length $K$ represents the information the receiver ultimately wants. The modulator maps the data word into its signal representation ($\mathbf{u}_\alpha \rightarrow \mathbf{s}_\alpha$) and transmits a continuous-time signal using an antipodal signal set. Viewed from the framework of information processing, we say that the information is encoded in the received signal vector $\mathbf{r}_\alpha$.

We calculate two KL distances. The first is between the distributions of the two received baseband signal vectors $\mathbf{r}_{\alpha_1}, \mathbf{r}_{\alpha_2}$ at the input of the detector. The second is between the discrete distributions associated with the corresponding output vectors $\mathbf{w}_{\alpha_1}, \mathbf{w}_{\alpha_2}$. If the detector makes hard decisions, each of the $K$ elements of $\mathbf{w}_\alpha$ belongs to the set $\{0, 1\}$ (one bit per decision). If the detector makes 2-bit soft decisions, then each element of $\mathbf{w}_\alpha$ belongs to the set $\mathcal{S} = \{00, 01, 10, 11\}$. We denote the input and output KL distances by $\mathcal{D}_\mathbf{r}(\alpha_1 \| \alpha_2)$ and $\mathcal{D}_\mathbf{w}(\alpha_1 \| \alpha_2)$, respectively. These distances quantify the informational change between the inputs and outputs of the detector.

Through Stein's Lemma [2], the KL distance is the exponential decay rate of the false alarm probability of an optimum Neyman-Pearson detector. Thus, $\mathcal{D}_\mathbf{r}(\alpha_1 \| \alpha_2)$ and $\mathcal{D}_\mathbf{w}(\alpha_1 \| \alpha_2)$ quantify our ability to discriminate between the two information bearing signals at the input and output of the detector. Because of the DPT [1], the detector can at best preserve the distance presented at their input and at worst, reduce it to zero causing the ultimate recipient of the transmission to lose all ability to discern the informational change.

The performance criterion we use is the *information transfer ratio*, denoted by $\gamma$, and defined as the ratio of the KL distances at the input and output of any system.

$$\gamma_{\text{det}} = \frac{\mathcal{D}_\mathbf{w}(\alpha_1 \| \alpha_2)}{\mathcal{D}_\mathbf{r}(\alpha_1 \| \alpha_2)} \tag{1}$$

It is a number between zero and one and reflects the fraction of the informational change preserved across a system. Ideally, the information transfer ratio across the detector would equal one indicating no informational loss. However in reality, we expect informational losses because the probability of error is never zero. Here, we constrast the performance of a detector making 2-bit soft decisions with one making hard decisions.

## 2. KULLBACK-LEIBLER DISTANCE CALCULATIONS

Each transmitted data word induces a probability distribution on the received signal vector at the output of the demodulator. For example, if the channel adds white Gaussian noise, $\mathbf{r}_\alpha$ would be a jointly normal random vector with mean vector $\pm\sqrt{E_b}\mathbf{1}$ and covariance $N_0/2\mathbf{I}_K$, where $\mathbf{1}$ is a vector of ones and $\mathbf{I}_K$ is a $K \times K$ identity matrix. ($E_b$ is the energy per data bit.) The statistical independence of the received vector elements allows us to write the KL distance at the input of the detector as a sum of the distances between each received vector element [3].

$$\mathcal{D}_\mathbf{r}(\alpha_1 \| \alpha_2) = \sum_{j=1}^{K} \mathcal{D}_{r_j}(\alpha_1 \| \alpha_2) \tag{2}$$

This work was supported by the National Science Foundation under Grant CCR-0105558.

[1]The word *distance* does not imply a metric since the KL distance is not symmetric in its arguments and does not satisfy the triangle inequality.

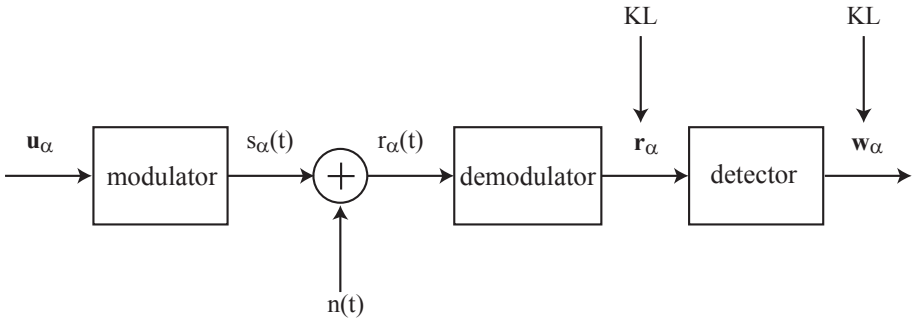


**Fig. 1**. Two binary data blocks $\mathbf{u}_{\alpha_1}$, $\mathbf{u}_{\alpha_2}$ are separately transmitted. The Kullback-Leibler distance between the distributions induced by each of the data blocks is calculated at the input and output of the detector. The ratios of the input and output distances provide a measure of how well the detector preserves the informational change encoded the input signals.

We can further simplify this expression because $\mathcal{D}_{r_j}(\alpha_1\|\alpha_2) = 0$ if the $j^{th}$ bits in each word are the same.

$$\mathcal{D}_{\mathbf{r}}(\alpha_1\|\alpha_2) = d_H(\mathbf{u}_{\alpha_1}, \mathbf{u}_{\alpha_2}) \cdot \mathcal{D}_r(\alpha_1\|\alpha_2) \tag{3}$$

Here, $d_H(\mathbf{u}_{\alpha_1}, \mathbf{u}_{\alpha_2})$ represents the Hamming distance between the datawords. Table 1 lists the KL distances $\mathcal{D}_r(\alpha_1\|\alpha_2)$ for various noise distributions as a function of *SNR*.

### 2.1. Hard Decisions

If the detector makes hard decisions it compares each received sample $r_{\alpha j}$ $(j = 1, \ldots, K)$ to a threshold and declares as its output either a one or a zero. The detected binary word $\mathbf{w}_\alpha$ is the collection of $K$ such outputs. We calculate the KL distance at the output of the detector by viewing each binary vector $\mathbf{w}_n$ $(n = 1, \ldots, 2^K)$ as the output of a binary symmetric channel with error probability $P_e$. (See Table 1 for expressions of $P_e$ for different noise distributions.) Accordingly, the probability of receiving $\mathbf{w}_n$ when we transmit $\mathbf{u}_\alpha$ is

$$\Pr[\mathbf{w}_n|\mathbf{u}_\alpha] = P_e^{d_H(\mathbf{w}_n,\mathbf{u}_\alpha)}(1-P_e)^{K-d_H(\mathbf{w}_n,\mathbf{u}_\alpha)}.$$

These probabilities define the discrete distribution over the output of the detector. Thus, by definition we obtain

$$\begin{aligned}\mathcal{D}_{\mathbf{w}}(\alpha_1\|\alpha_2) &= \sum_{j=1}^{K}\mathcal{D}_{w_j}(\alpha_1\|\alpha_2) \\ &= \sum_{n=1}^{2^K}\Pr[\mathbf{w}_n|\mathbf{u}_{\alpha_1}]\log\frac{\Pr[\mathbf{w}_n|\mathbf{u}_{\alpha_1}]}{\Pr[\mathbf{w}_n|\mathbf{u}_{\alpha_2}]}.\end{aligned} \tag{4}$$

Like equation (2), we can simplify equation (4) because $\mathcal{D}_{w_j}(\alpha_1\|\alpha_2) = 0$ if the $j^{th}$ bits in each word are the same.

$$\mathcal{D}_{\mathbf{w}}(\alpha_1\|\alpha_2) = d_H(\mathbf{u}_{\alpha_1}, \mathbf{u}_{\alpha_2}) \cdot \left[(1-P_e)\log\frac{1-P_e}{P_e} + P_e\log\frac{P_e}{1-P_e}\right] \tag{5}$$

The bracketed term is the KL distance between two binary distributions which result from the transmission of corresponding bits of $\mathbf{u}_{\alpha_1}$ and $\mathbf{u}_{\alpha_2}$.

### 2.2. 2-Bit Soft Decisions

When the detector makes 2-bit soft decisions, each element $w_j$ is distributed over the set $\mathcal{S}$ conditioned on a transmitted bit. As depicted in Figure 2, these probabilities are the probabilities of the four different regions defined by the thresholds 0 and $\pm\theta$.

$$\Pr[w_j = n|u_j = 0] = \int_{\mathcal{R}_n} p_{r_j|u_j=0}(x)\,dx$$

For soft decisions, $n \in \mathcal{S}$. The probabilities given $u_j = 1$ are defined similarly with the distribution $p_{r_j|u_j=1}$. In the next section, we find the thresholds that maximize $\gamma$ for specified *SNR* values subject to the following constraints. First, we require zero to always be a threshold. Second, we require the remaining two thresholds be symmetric about zero. The first limitation seems reasonable because of the inherent symmetry of the problem. We impose the second to make the maximization tractable. As before, because each bit is transmitted independently, the total KL distance between the distributions of $\mathbf{w}_{\alpha_1}, \mathbf{w}_{\alpha_2}$ is again the sum of the distances between each element.

$$\mathcal{D}_{\mathbf{w}}(\alpha_1\|\alpha_2) = \sum_{j=1}^{K}\mathcal{D}_{w_j}(\alpha_1\|\alpha_2)$$

With the imposition of the above restrictions we can write

$$\mathcal{D}_{\mathbf{w}}(\alpha_1\|\alpha_2) = d_H(\mathbf{u}_{\alpha_1}, \mathbf{u}_{\alpha_2}) \cdot \sum_{n=0}^{3}\Pr[n|u_j=0]\log\frac{\Pr[n|u_j=0]}{\Pr[n|u_j=1]} \tag{6}$$

where the value of $n$ corresponds to the decimal equivalent of the binary numbers in $\mathcal{S}$.

## 3. RESULTS

### 3.1. Hard Decisions

From equations (3) and (5) we immediately notice that the information transfer ratio across the detector is independent of the input data words. In particular, this means the performance of hard decision detectors is invariant to data word length and to the Hamming distance between the two input data words $\mathbf{u}_{\alpha_1}, \mathbf{u}_{\alpha_2}$. Furthermore, in this setting the KL distances are symmetric

| *Noise Distribution* | $\mathcal{D}_r\left(\alpha_1 \| \alpha_2\right)$ | $P_e$ | $\gamma$ $SNR \to 0$ | $\gamma$ $SNR \to \infty$ |
|---|---|---|---|---|
| Gaussian | $4\xi$ | $Q\left(\sqrt{2\xi}\right)$ | $\frac{2}{\pi}$ | $\frac{1}{4}$ |
| Laplacian | $e^{-4\sqrt{\xi}} - 1 + 4\sqrt{\xi}$ | $\frac{1}{2}\, e^{-2\sqrt{\xi}}$ | 1 | $\frac{1}{2}$ |
| Hyperbolic Secant | $-2\ln\left[\text{sech}\left(\frac{\pi}{2}\sqrt{2\xi}\right)\right]$ | $\frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left[\sinh\left(\frac{\pi}{2}\sqrt{2\xi}\right)\right]$ | $\frac{8}{\pi 2}$ | $\frac{1}{2}$ |
| Cauchy | $\ln\left(1+2\xi\right)$ | $\frac{1}{2} - \frac{1}{\pi}\,\tan^{-1}\left(\sqrt{\xi}\right)$ | $\frac{8}{\pi 2}$ | $\frac{1}{2}$ |

**Table 1**. The Kullback-Leibler distances between the received random variables $r_{\alpha_1 j}$ and $r_{\alpha_2 j}$ and the detector's hard decision bit error probabilities are shown in columns two and three for various noise distributions. In each expression $\xi = E_b/N_0$ is the signal-to-noise ratio per bit (*SNR*). For the Cauchy distribution, the quantity $N_0$ is understood to be the "width" parameter. The fourth and fifth columns list the asymptotic values of the information transfer ratio across when the detector makes hard decisions.

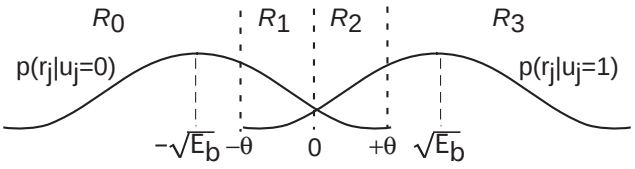


**Fig. 2**. The probabilities over $\mathcal{S}$ for the 2-bit soft decision detector are simply the probability mass of $p_{r_j|u_j}$ over the appropriate regions $\mathcal{R}_n, n \in \mathcal{S}$. There are three thresholds $0, \pm\theta$, but because of the symmetry of the problem we can maximize the information transfer ratio as a function of $\theta$ to find its optimal value for a specified *SNR*.

($\mathcal{D}_{\mathbf{r}}\left(\alpha_1 \| \alpha_2\right) = \mathcal{D}_{\mathbf{r}}\left(\alpha_2 \| \alpha_1\right)$ and $\mathcal{D}_{\mathbf{w}}\left(\alpha_1 \| \alpha_2\right) = \mathcal{D}_{\mathbf{w}}\left(\alpha_2 \| \alpha_1\right)$). We plot information transfer ratios for four noise distributions as a function of *SNR* in Figure 3. Table 1 lists their asymptotic values. *These curves show the informational loss for making hard decisions.* Notice the decrease in performance as the *SNR* increases. This means hard decision detectors better preserve informational changes at lower *SNR* values than at higher values. In other words, hard decision detectors are more sensitive to bit changes at lower *SNR* values than at higher values. *However*, even though the detector is less informationally efficient with *SNR*, the loss is not great compared to other physical systems (e.g. neurons [4]). We prove in Appendix A that likelihood ratio tests maximize the information transfer ratio across binary detectors. Thus, the curves in Figure 3 represent the best achievable performance across *any* hard decision detector.

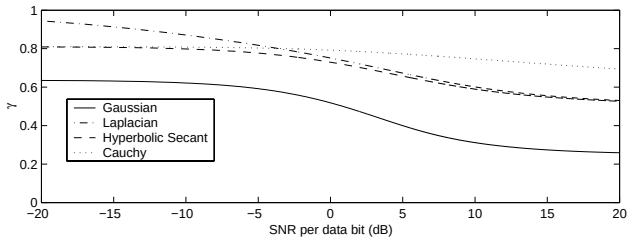


**Fig. 3**. The information transfer ratio quantifies the informational loss associated with making hard decisions. At lower SNR values likelihood ratio detectors are more sensitive to informational changes than at higher SNR values.

### 3.2. 2-Bit Soft Decisions

Our immediate goal is to find the threshold $\theta$ which maximizes $\gamma$ for a specified *SNR*.

$$\max_{\theta>0} \gamma(\theta) = \frac{\mathcal{D}_{\mathbf{w}}\left(\alpha_1 \| \alpha_2\right)}{\mathcal{D}_{\mathbf{r}}\left(\alpha_1 \| \alpha_2\right)} \tag{7}$$

Because the denominator does not depend upon $\theta$, maximizing $\mathcal{D}_{\mathbf{w}}\left(\alpha_1 \| \alpha_2\right)$ is equivalent to maximizing $\gamma$. For the four distributions listed in Table 1 it can be easily shown that this maximization is not a convex optimization problem; however, at least one maximum does indeed exist. We numerically computed a maximizing $\theta$ for several values of *SNR* using Matlab's optimization program `fmincon`. They are listed in Figure 4 along with the associated information transfer ratio plots.

These plots clearly show that with the addition of just one bit of soft decision, significant informational gains can be realized over hard decision detectors. Not surprisingly, the gains are not uniform but are centered around the *SNR* value specified in the maximization of $\theta$. The performance of hard decision detectors (Figure 3) serves as a lower bound for all 2-bit soft decision schemes. This fact becomes evident by direct application of the DPT. Because hard decisions are a special case of 2-bit soft decisions we can model the hard decision detector as a 2-bit soft decision detector cascaded with another system. Thus, the overall informational loss ($\gamma$ over both systems) must be greater than or equal to the loss across the soft decision detector [1].

## 4. CONCLUSION

By applying the precepts of information processing, we were able to take new look at an old problem. In fact there is nothing new in the notion that soft decisions provide benefits over hard decisions, but what is new is our analysis approach and the interpretation of the results. Quantifying performance in terms of the information transfer ratio allows us to tackle non-Gaussian noise environments and analyze detector performance in a new information-theoretic sense. Our results also suggest that this method has the potential to determine how soft (i.e. how many bits) a decision has to be in order to achieve a specified performance level.

## A. APPENDIX

Consider a general binary detection problem where $\mathbf{r}_{\alpha_1}$ and $\mathbf{r}_{\alpha_2}$ are two possible received signal vectors presented at the input of

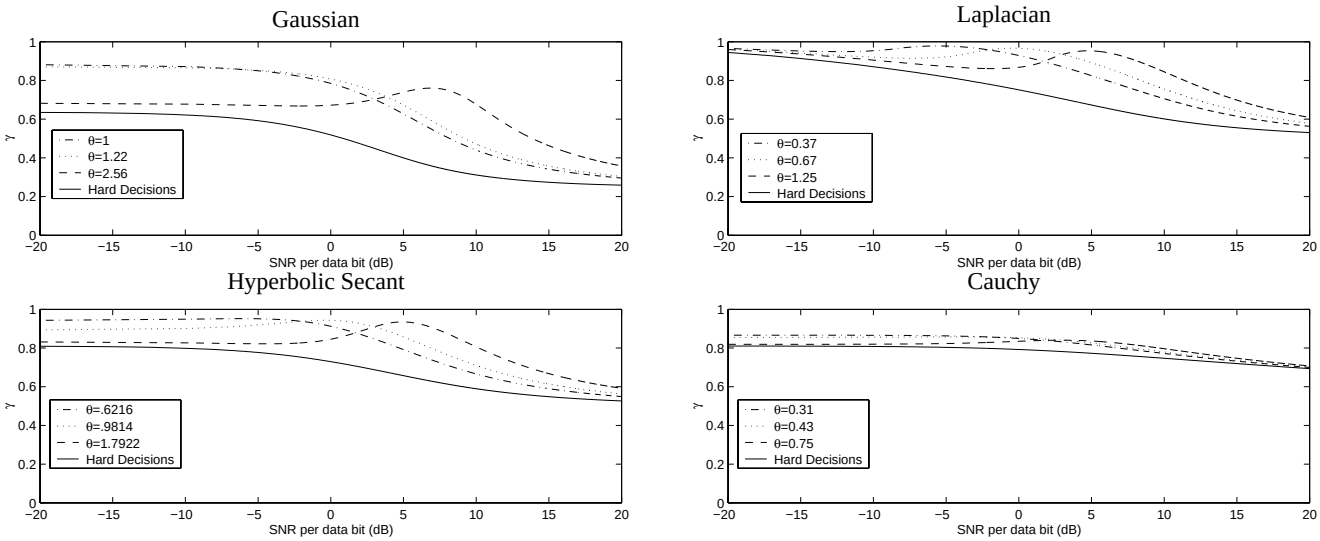


**Fig. 4**. Each plot compares hard and soft detector performance in terms of the information transfer ratio in various noise environments. The three threshold values listed in each case are the thresholds which maximize $\gamma$ at *SNR* values of $-5, 0, 5$ dB, respectively. The solid black curve is the performance of the hard decision detector. All curves were generated with input data words $u_{\alpha_1} = 0000, u_{\alpha_2} = 1011$. For the Laplacian and Cauchy cases, the parameter $\sigma = 1$ (noise variance for Laplacian and width parameter for Cauchy).

the detector under hypothesis $\alpha_1$ and $\alpha_2$ respectively. Let $p(\mathbf{r}|\alpha_1)$ and $p(\mathbf{r}|\alpha_2)$ be conditional probability density functions associated with each hypothesis. Denote the output decisions of the detector as $\Lambda_1$ and $\Lambda_2$.

The information transfer ratio equals

$$\gamma = \frac{\mathcal{D}_\Lambda\left(\alpha_1 \| \alpha_2\right)}{\mathcal{D}_{\mathbf{r}}\left(\alpha_1 \| \alpha_2\right)} = \frac{P_D \log\left(P_D/P_F\right) + \left(1-P_D\right)\log\left(1-P_D\right)/\left(1-P_F\right)}{\int p(\mathbf{r}|\alpha_1) \log \frac{p(\mathbf{r}|\alpha_1)}{p(\mathbf{r}|\alpha_2)}\, d\mathbf{r}}$$

where $P_D$ is the probability of detection and $P_F$ is the probability of false alarm. Explicitly,

$$p(\Lambda_1|\alpha_1) = 1 - P_F \qquad p(\Lambda_2|\alpha_1) = P_F$$
$$p(\Lambda_1|\alpha_2) = 1 - P_D \qquad p(\Lambda_2|\alpha_2) = P_D.$$

Maximizing $\gamma$ is equivalent to maximizing the numerator which translates into finding values of $P_D$ and $P_F$ which maximize

$$P_D \log\left(\frac{P_D}{P_F}\right) + \left(1-P_D\right)\log\left(\frac{1-P_D}{1-P_F}\right) = -H(P_D) - P_D \log P_F - \left(1-P_D\right)\log\left(1-P_F\right). \qquad (8)$$

where $H(\cdot)$ denotes the entropy function of a Bernoulli distribution. Since $P_D$ and $P_F$ are coupled they can not be independently optimized, so without loss of generality, assume $P_F = a$ and $P_D = a + l$. Substituting these values into equation (8) and setting its derivative equal to zero we obtain

$$\log\left[\frac{-(a2 + al) + a + l}{-(a2 + al) + a}\right] = 0.$$

For a given value of $a$ ($P_F$), we note that the derivative is positive for $l > 0$, negative for $l < 0$, and zero when $l = 0$ (minimum). Thus to maximize the numerator of equation (8) we choose the largest possible $l$ but constrained to $0 \le l \le 1-a$. The upper bound results from the fact that $P_D$ and $P_F$ are probabilities and thus must be between zero and one.

Formally, for a given false-alarm probability

$$\max_{l<1-a} l = \max_{P_D} P_D - P_F = \max_{\Lambda_1} \int_{\Lambda_1} p(\mathbf{r}|\alpha_2) - p(\mathbf{r}|\alpha_1)\, d\mathbf{r}$$

Therefore $\Lambda_1$ should be defined as

$$\Lambda_1 = \{\mathbf{r} | p(\mathbf{r}|\alpha_2) > p(\mathbf{r}|\alpha_1)\}$$

which is exactly the condition of the likelihood ratio test. This result is general and holds for all noise distributions. □

## B. REFERENCES


[1] S. Sinanović and D. H. Johnson, "Toward a theory of information processing," *Submitted to IEEE Trans on Signal Processing*, Jun 2002.

[2] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, 1991.

[3] Sinan Sinanović, *Toward a Theory of Information Processing*, 1999, Master of Science Thesis, Rice University, Houston TX.

[4] C. Rozell, D.H. Johnson, and R.M. Glantz, "Information processing during transient responses in the crayfish visual system," *Computational Neuroscience '02*, vol. Chicago, IL, 2002.